Phylogenetics in the Genomic Era

Celine Scornavacca, Frédéric Delsuc, Nicolas Galtier

Phylogenetics in the Genomic Era

Edited by
Celine Scornavacca
Frédéric Delsuc
Nicolas Galtier

Editors

Celine Scornavacca
Institut des Sciences de l'Évolution Université de Montpellier, CNRS, IRD, EPHE
Place Eugène Bataillon 34095 Montpellier Cedex 05, France
celine.scornavacca@umontpellier.fr

Frédéric Delsuc
Institut des Sciences de l'Évolution Université de Montpellier, CNRS, IRD, EPHE
Place Eugène Bataillon 34095 Montpellier Cedex 05, France
frederic.delsuc@umontpellier.fr

Nicolas Galtier
Institut des Sciences de l'Évolution Université de Montpellier, CNRS, IRD, EPHE
Place Eugène Bataillon 34095 Montpellier Cedex 05, France
nicolas.galtier@umontpellier.fr

Hosted by
Hyper Articles en Ligne (HAL): https://hal.inria.fr.

Digital Object Identifier: https://hal.inria.fr/PGE
International Standard Book Number: 978-2-9575069-0-3

Publication date
First version published on April 15, 2020. Current version (v2) published on November 25, 2020.

Credits
The book has been edited using a customized version of the `OASIcs Class`, licensed under the term of the LPPL (Version 1.3c, https://www.latex-project.org/lppl.txt), available from https://github.com/dagstuhl-publishing/styles.

The "scissors cut dollar paper bill" icon has been obtained by modifying the "debt" Creative Commons Universal 1.0 icon (CC0 1.0, https://creativecommons.org/licenses/by/1.0/) available from https://search.creativecommons.org/photos/b68e4b13-7348-42c6-99fb-4bf6002fbf74.

The almond tree picture on the cover page is one of the beautiful Maël Bathfield's panoramas (copyright Maël Bathfield, https://photo.maelbathfield.net).

This book has been completely handled by researchers. No publisher has been paid.

To Open Science

Contents

Part 1: Phylogenetic analyses in the genomic era

Chapter 1.1 A gentle Introduction to Probabilistic Evolutionary Models
Tal Pupko and Itay Mayrose 1.1:1–1.1:21

Chapter 1.2 Efficient Maximum Likelihood Tree Building Methods
Alexandros Stamatakis and Alexey M. Kozlov 1.2:1–1.2:18

Chapter 1.3 Using RAxML-NG in Practice
Alexey M. Kozlov and Alexandros Stamatakis 1.3:1–1.3:25

Chapter 1.4 The Bayesian Approach to Molecular Phylogeny
Nicolas Lartillot 1.4:1–1.4:17

Chapter 1.5 PhyloBayes: Bayesian Phylogenetics Using Site-heterogeneous Models
Nicolas Lartillot 1.5:1–1.5:16

Part 2: Data quality, model adequacy

Chapter 2.1 To What Extent Current Limits of Phylogenomics Can Be Overcome?
Paul Simion, Frédéric Delsuc, and Hervé Philippe 2.1:1–2.1:34

Chapter 2.2 Strengths and Limits of Multiple Sequence Alignment and Filtering Methods
Vincent Ranwez and Nathalie Chantret 2.2:1–2.2:36

Chapter 2.3 Accurate Alignment of (Meta)barcoding Data Sets using MACSE
Frédéric Delsuc and Vincent Ranwez 2.3:1–2.3:31

Chapter 2.4 Orthology: Definitions, Prediction, and Impact on Species Phylogeny Inference
Rosa Fernández, Toni Gabaldón, and Christophe Dessimoz 2.4:1–2.4:14

Chapter 2.5 Ancestral Genome Organization as a Diagnosis Tool for Phylogenomics
Eric Tannier, Adelme Bazin, Adrián A. Davín, Laurent Guéguen, Sèverine Bérard, and Cédric Chauve 2.5:1–2.5:19

Part 3: Resolving phylogenomic conflicts

Chapter 3.1 The Sources of Phylogenetic Conflicts
Dominik Schrempf and Gergely Szöllősi 3.1:1–3.1:23

Chapter 3.2 Reconciling Gene Trees with Species Trees
Bastien Boussau and Celine Scornavacca 3.2:1–3.2:23

Chapter 3.3 The Multi-species Coalescent Model and Species Tree Inference
Bruce Rannala, Scott V. Edwards, Adam Leaché, and Ziheng Yang 3.3:1–3.3:21

Phylogenetics in the genomic era.
Editor: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier
A book completely handled by researchers.
No publisher has been paid.

Chapter 3.4 The Concatenation Question
David Bryant and Matthew W. Hahn 3.4:1–3.4:23

Part 4: Functional evolutionary genomics

Chapter 4.1 Phylogenomics and Genome Annotation
Anamaria Necsulea 4.1:1–4.1:26

Chapter 4.2 Molecular Evolution and Gene Function
Marc Robinson-Rechavi 4.2:1–4.2:20

Chapter 4.3 The Expression Comparison Tool in Bgee
Marc Robinson-Rechavi, Valentine Rech de Laval, Frédéric B. Bastian, Julien Wollbrett, Bgee Team 4.3:1–4.3:4

Chapter 4.4 Substitution Rate Analysis and Molecular Evolution
Lindell Bromham 4.4:1–4.4:21

Chapter 4.5 Detecting Adaptation from Multi-species Protein-coding DNA Sequence Alignments
Christine Lowe and Nicolas Rodrigue 4.5:1–4.5:18

Chapter 4.6 The Nature and Phylogenomic Impact of Sequence Convergence
Zhengting Zou and Jianzhi Zhang 4.6:1–4.6:17

Part 5: Phylogenomic applications

Chapter 5.1 Inferring the Timescale of Phylogenetic Trees from Fossil Data
Walker Pett and Tracy A. Heath 5.1:1–5.1:18

Chapter 5.2 Estimating a Time-calibrated Phylogeny of Fossil and Extant Taxa using RevBayes
Joëlle Barido-Sottani, Joshua A. Justison, April M. Wright, Rachel C. M. Warnock, Walker Pett, and Tracy A. Heath 5.2:1–5.2:23

Chapter 5.3 Efficiently Analysing Large Viral Data Sets in Computational Phylogenomics
Anna Zhukova, Olivier Gascuel, Sebastián Duchêne, Daniel L. Ayres, Philippe Lemey, and Guy Baele 5.3:1–5.3:43

Chapter 5.4 BEAGLE 3 High-performance Computational Library for Phylogenetic Inference
Daniel L. Ayres, Philippe Lemey, Guy Baele, and Marc A. Suchard 5.4:1–5.4:9

Chapter 5.5 Species Delimitation
Bruce Rannala and Ziheng Yang 5.5:1–5.5:18

Chapter 5.6 A Tutorial on the Use of BPP for Species Tree Estimation and Species Delimitation
Tomáš Flouri, Bruce Rannala, and Ziheng Yang 5.6:1–5.6:16

Preface

A note to the reader. This journey in the collaborative publishing has started with the unreasonable request from a well-known publisher to make us pay 20,000 euros for an open access book that we (the editors), would have had to handle from A to Z (choose the chapters, find the authors, get the contracts signed, format the manuscript etc ...). It did not seem right, so we asked the colleagues that already had shown interest in writing a chapter for our book, to trust us. They did, and with the help of the HAL team and Christine Bibal, and some knowledge of LaTeX, here we are. Proud to present you our collective effort, licensed under CC BY-NC-ND 4.0. Enjoy it freely.

Molecular phylogenetics was born in the middle of the 20$^{\text{th}}$ century, when the advent of protein and DNA sequencing offered a novel way to study the evolutionary relationships between living organisms. The first 50 years of the discipline can be seen as a long quest for resolving power. The goal – reconstructing the tree of life – seemed to be unreachable, the methods were heavily debated, and the data limiting. Maybe for these reasons, even the relevance of the whole approach was repeatedly questioned, as part of the so-called molecules versus morphology debate. Controversies often crystalized around long-standing conundrums, such as the origin of land plants, the diversification of placental mammals, or the prokaryote/eukaryote divide. Some of these questions were resolved as gene and species samples increased in size. Over the years, molecular phylogenetics has gradually evolved from a brilliant, revolutionary idea to a mature research field centred on the problem of reliably building trees.

This logical progression was abruptly interrupted in the late 2000s. High-throughput sequencing arose and the field suddenly moved into something entirely different. Access to genome-scale data profoundly reshaped the methodological challenges, while opening an amazing range of new application perspectives. Phylogenetics left the realm of systematics to occupy a central place in one of the most exciting research fields of this century – genomics. This is what this book is about: how we do trees, and what we do with trees, in the current phylogenomic era.

One obvious, practical consequence of the transition to genome-scale data is that the most widely used tree-building methods, which are based on probabilistic models of sequence evolution, require intensive algorithmic optimization to be applicable to current datasets. This problem is considered in Part 1 of the book, which includes a general introduction to Markov models (Chapter 1.1) and a detailed description of how to optimally design and implement Maximum Likelihood (Chapter 1.2) and Bayesian (Chapter 1.4) phylogenetic inference methods. The importance of the computational aspects of modern phylogenomics is such that efficient software development is a major activity of numerous research groups in the field. We acknowledge this and have included seven "How to" chapters presenting recent updates of major phylogenomic tools – RAxML (Chapter 1.3), PhyloBayes (Chapter 1.5), MACSE (Chapter 2.3), Bgee (Chapter 4.3), RevBayes (Chapter 5.2), Beagle (Chapter 5.4), and BPP (Chapter 5.6).

Genome-scale data sets are so large that statistical power, which had been the main limiting factor of phylogenetic inference during previous decades, is no longer a major issue. Massive data sets instead tend to amplify the signal they deliver – be it biological or artefactual – so that bias and inconsistency, instead of sampling variance, are the main problems with phylogenetic inference in the genomic era. Part 2 covers the issues of data

Phylogenetics in the genomic era.
Editor: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier
A book completely handled by researchers.
No publisher has been paid.

quality and model adequacy in phylogenomics. Chapter 2.1 provides an overview of current practice and makes recommendations on how to avoid the more common biases. Two chapters review the challenges and limitations of two key steps of phylogenomic analysis pipelines, sequence alignment (Chapter 2.2) and orthology prediction (Chapter 2.4), which largely determine the reliability of downstream inferences. The performance of tree building methods is also the subject of Chapter 2.5, in which a new approach is introduced to assess the quality of gene trees based on their ability to correctly predict ancestral gene order.

Analyses of multiple genes typically recover multiple, distinct trees. Maybe the biggest conceptual advance induced by the phylogenetic to phylogenomic transition is the suggestion that one should not simply aim to reconstruct "the" species tree, but rather be prepared to make sense of forests of gene trees. Chapter 3.1 reviews the numerous reasons why gene trees can differ from each other and from the species tree, and what the implications are for phylogenetic inference. Chapter 3.2 focuses on gene trees/species trees reconciliation methods that account for gene duplication/loss and horizontal gene transfer among lineages. Incomplete lineage sorting is another major source of phylogenetic incongruence among loci, which recently gained attention and is covered by Chapter 3.3. Chapter 3.4 concludes this part by taking a user's perspective and examining the pros and cons of concatenation versus separate analysis of gene sequence alignments.

Modern genomics is comparative and phylogenetic methods are key to a wide range of questions and analyses relevant to the study of molecular evolution. This is covered by Part 4. We argue that genome annotation, either structural or functional, can only be properly achieved in a phylogenetic context. Chapters 4.1 and 4.2 review the power of these approaches and their connections with the study of gene function. Molecular substitution rates play a key role in our understanding of the prevalence of nearly neutral versus adaptive molecular evolution, and the influence of species traits on genome dynamics (Chapter 4.4). The analysis of substitution rates, and particularly the detection of positive selection, requires sophisticated methods and models of coding sequence evolution (Chapter 4.5). Phylogenomics also offers a unique opportunity to explore evolutionary convergence at a molecular level, thus addressing the long-standing question of predictability versus contingency in evolution (Chapter 4.6).

The development of phylogenomics, as reviewed in Parts 1 through 4, has resulted in a powerful conceptual and methodological corpus, which is often reused for addressing problems of interest to biologists from other fields. Part 5 illustrates this application potential via three selected examples. Chapter 5.1 addresses the link between phylogenomics and palaeontology; i.e., how to optimally combine molecular and fossil data for estimating divergence times. Chapter 5.3 emphasizes the importance of the phylogenomic approach in virology and its potential to trace the origin and spread of infectious diseases in space and time. Finally, Chapter 5.5 recalls why phylogenomic methods and the multi-species coalescent model are key in addressing the problem of species delimitation – one of the major goals of taxonomy.

It is hard to predict where phylogenomics as a discipline will stand in even 10 years. Maybe a novel technological revolution will bring it to yet another level? We strongly believe, however, that tree thinking will remain pivotal in the treatment and interpretation of the deluge of genomic data to come. Perhaps a prefiguration of the future of our field is provided by the daily monitoring of the current Covid-19 outbreak via the phylogenetic analysis of coronavirus genomic data in quasi real time – a topic of major societal importance, contemporary to the publication of this book, in which phylogenomics is instrumental in helping to fight disease.

The digital version of the book, and of its individual chapters, are accessible for free

 The editors would like to thank all the authors for their tremendous contributions and enthusiastic participation to this collective effort. We are grateful to David Manley, Christine Bibal and the HAL platform for their help in editing/distributing this book.

List of Authors (in order of appearance)

Tal Pupko
School of Molecular Cell Biology & Biotechnology, Tel Aviv University
Tel-Aviv, Israel
TalP@tauex.tau.ac.il

Itay Mayrose
School of Plant Sciences and Food security, Tel Aviv University
Tel-Aviv, Israel
itaymay@post.tau.ac.il

Alexandros Stamatakis
Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies
Karlsruhe Institute of Technology, Institute for Theoretical Informatics
Heidelberg, Germany
Alexandros.Stamatakis@h-its.org

Alexey M. Kozlov
Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies
Heidelberg, Germany
Alexey.Kozlov@h-its.org

Nicolas Lartillot
Laboratoire de Biométrie et Biologie Evolutive (LBBE), CNRS, Université de Lyon
Villeurbanne, France
nicolas.lartillot@univ-lyon1.fr

Paul Simion
Laboratoire d'Ecologie et Génétique Evolutive (LEGE), URBE, University of Namur
Namur, Belgium
polo.simion@gmail.com

Frédéric Delsuc
Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier
Montpellier, France
frederic.delsuc@umontpellier

Hervé Philippe
Station d'Ecologie Théorique et Expérimentale, UMR CNRS 5321
Moulis, France
herve.philippe@sete.cnrs.fr

Vincent Ranwez
AGAP, CIRAD, INRA, Montpellier SupAgro, Université de Montpellier
Montpellier, France
vincent.ranwez@supagro.fr

Nathalie Chantret
AGAP, CIRAD, INRA, Montpellier SupAgro, Université de Montpellier
Montpellier, France
nathalie.chantret@inra.fr

Rosa Fernández
Institute of Evolutionary Biology, Spanish National Research Council, University Pompeu Fabra
Barcelona, Spain
rosa.fernandez@ibe.upf-csic.es

Toni Gabaldon
Barcelona Supercomputing Centre
Institute for Research in Biomedicine
The Barcelona Institute of Science and Technology
Catalan Institution for Research and Advanced Studies
Barcelona, Spain
toni.gabaldon.bcn@gmail.com

Christophe Dessimoz
Department of Computational Biology & Center for Integrative Genomics, University of Lausanne
Lausanne, Switzerland
Swiss Institute of Bioinformatics
Lausanne, Switzerland
Centre for Life's Origins and Evolution & Department of Computer Science
University College London, UK
Christophe.Dessimoz@unil.ch

Phylogenetics in the genomic era.
Editor: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier
A book completely handled by researchers.
No publisher has been paid.

Eric Tannier
Laboratoire de Biométrie et Biologie Evolutive, UMR5558 CNRS, INRIA, Université Lyon 1
Villeurbanne, France
eric.tannier@inria.fr

Adelme Bazin
Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université d'Evry, Université Paris-Saclay
Evry, France
adelme.bazin@genoscope.cns.fr

Adrián A. Davín
RIKEN Center for Advanced Intelligence Project (AIP)
Kyoto, Japan
adrian.arellanodavin@riken.jp

Laurent Guéguen
Laboratoire de Biométrie et Biologie Evolutive, UMR5558 CNRS, Université de Lyon
Villeurbanne, France
laurent.gueguen@univ-lyon1.fr

Sèverine Bérard
Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier
Montpellier, France
severine.berard@univ-montellier.fr

Cédric Chauve
Department of Mathematics, Simon Fraser University
Burnaby, BC, Canada
LaBRI, Université de Bordeaux
Talence, France
cedric.chauve@sfu.ca

Dominik Schrempf
Department of Biological Physics, Eötvös University
Budapest, Hungary

Gergely J. Szöllősi
ELTE-MTA Lendület Evolutionary Genomics Research Group
Budapest, Hungary
Department of Biological Physics, Eötvös University
Budapest, Hungary
Evolutionary Systems Research Group, Centre for Ecological Research, Hungarian Academy of Sciences
Tihany, Hungary

Bastien Boussau
Laboratoire de Biométrie et Biologie Évolutive (LBBE), CNRS, Université de Lyon
Villeurbanne, France
bastien.boussau@univ-lyon1.fr

Celine Scornavacca
Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier
Montpellier, France
celine.scornavacca@umontpellier.fr

Bruce Rannala
Department of Evolution and Ecology, University of California Davis
Davis, CA, USA
brannala@ucdavis.edu

Scott V. Edwards
Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University
Cambridge, MA, USA
sedwards@fas.harvard.edu

Adam Leaché
Department of Biology & Burke Museum of Natural History and Culture, University of Washington
Seattle, WA, USA
leache@uw.edu

Ziheng Yang
Department of Genetics, Evolution and Environment, University College London
London, UK
z.yang@ucl.ac.uk

David Bryant
Department of Mathematics and Statistics, University of Otago Dunedin, New Zealand
david.bryant@otago.ac.nz

Matthew W. Hahn
Department of Biology, Department of Computer Science, Indiana University
Bloomington IN, USA
mwh@indiana.edu

Anamaria Necsulea
Laboratoire de Biométrie et Biologie Évolutive, CNRS, Université de Lyon, Université Lyon 1
Villeurbanne, France
anamaria.necsulea@univ-lyon1.fr

Marc Robinson-Rechavi
Department of Ecology and Evolution, University of Lausanne
Lausanne, Switzerland
Swiss Institute of Bioinformatics
Lausanne, Switzerland
marc.robinson-rechavi@unil.ch

Valentine Rech de Laval
Department of Ecology and Evolution, University of Lausanne
Lausanne, Switzerland
Swiss Institute of Bioinformatics
Lausanne, Switzerland
valentine.rechdelaval@unil.ch

Frédéric B. Bastian
Department of Ecology and Evolution, University of Lausanne
Lausanne, Switzerland
Swiss Institute of Bioinformatics
Lausanne, Switzerland
frederic.bastian@unil.ch

Julien Wollbrett
Department of Ecology and Evolution, University of Lausanne
Lausanne, Switzerland
Swiss Institute of Bioinformatics
Lausanne, Switzerland
julien.wollbrett@unil.ch

Lindell Bromham
Macroevolution & Macroecology, Division of Ecology & Evolution, Research School of Biology, Australian National University
Canberra, Australia
lindell.bromham@anu.edu.au

Christine Lowe
Biological Informatics Centre of Excellence, Agriculture and Agri-Food Canada
Ottawa, Canada
christine.lowe@canada.ca

Nicolas Rodrigue
Department of Biology, Institute of Biochemistry, and School of Mathematics and Statistics, Carleton University
Ottawa, Canada
nicolas.rodrigue@carleton.ca

Zhengting Zou
Department of Ecology and Evolutionary Biology, University of Michigan
Ann Arbor, Michigan, USA
ztzou@umich.edu

Jianzhi Zhang
Department of Ecology and Evolutionary Biology, University of Michigan
Ann Arbor, Michigan, USA
jianzhi@umich.edu

Walker Pett
Department of Ecology, Evolution, & Organismal Biology, Iowa State University
Ames, Iowa, USA
willpett@iastate.edu

Tracy A. Heath
Department of Ecology, Evolution, & Organismal Biology, Iowa State University
Ames, Iowa, USA
phylo@iastate.edu

Joëlle Barido-Sottani
Department of Ecology, Evolution, & Organismal Biology, Iowa State University
Ames, Iowa, USA
joellebs@iastate.edu

Joshua A. Justison
Department of Ecology, Evolution, & Organismal Biology, Iowa State University
Ames, Iowa, USA
justison@iastate.edu

April M. Wright
Department of Biological Sciences, Southeastern Louisiana University
Hammond, LA 70402 USA
april.wright@selu.edu

Rachel C. M. Warnock
Technische Hochschule Zürich, Swiss Institute of Bioinformatics (SIB)
4058 Basel, Switzerland
rachel.warnock@bsse.ethz.ch

Anna Zhukova
Unité Bioinformatique Evolutive, Institut Pasteur
Paris, France
anna.zhukova@pasteur.fr

Olivier Gascuel
Unité Bioinformatique Evolutive, Institut Pasteur
Paris, France
olivier.gascuel@pasteur.fr

Sebastián Duchêne
Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, University of Melbourne
Melbourne, Australia
sebastian.duchene@unimelb.edu.au

Daniel L. Ayres
Center for Bioinformatics and Computational Biology, University of Maryland, USA
College Park, Maryland, USA
ayres@umiacs.umd.edu

Philippe Lemey
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven – University of Leuven
Leuven, Belgium
philippe.lemey@kuleuven.be

Guy Baele
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven – University of Leuven
Leuven, Belgium
guy.baele@kuleuven.be

Marc A. Suchard
Department of Biomathematics, Department of Biostatistics, Department of Human Genetics, University of California
Los Angeles, California, USA
msuchard@ucla.edu

Tomàs Flouri
Department of Genetics, Evolution and Environment, University College London
London, UK
t.flouris@ucl.ac.uk

Chapter 1.1 A gentle Introduction to Probabilistic Evolutionary Models

Tal Pupko
School of Molecular Cell Biology & Biotechnology, Tel Aviv University, Ramat Aviv, Tel-Aviv 69978, Israel
TalP@tauex.tau.ac.il

Itay Mayrose
School of Plant Sciences and Food security, Tel Aviv University, Ramat Aviv, Tel-Aviv 69978, Israel
itaymay@post.tau.ac.il

Abstract

A large body of research is dedicated to model sequence evolutionary dynamics. The evolutionary process may vary within groups of genes, among sites within a gene, between populations and among diverged species. Evolutionary models aiming to describe these dynamics must account for base pair substitutions as well as insertion and deletion (indel) events. Here, we explain the fundamental of continuous time Markov models used to describe sequence evolution. We begin by describing discrete Markov models, and slowly progress towards more realistic and more computationally complicated continuous time Markov models. Among other topics, we discuss nucleotide, amino acid, and codon models, among site rate variation, model reversibility, stationary distributions, rate matrix normalization, mixture models, indel models, and models of gene family evolution. Understanding the concepts presented here is vital for various phylogenomics analyses such as the inference of positive selection, alignment and phylogeny reconstruction, ancestral sequence reconstruction, and molecular dating.

Keywords and phrases A

How to cite: Tal Pupko and Itay Mayrose (2020). A gentle Introduction to Probabilistic Evolutionary Models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 1.1, pp. 1.1:1–1.1:21. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 General outline – the great importance of evolutionary models

The past decade has witnessed a revolution in evolutionary biology, driven by advances in high-throughput sequencing, functional genomics, and computational technologies. The genomic revolution opened up possibilities for rapidly generating large-scale sequencing data from non-model organisms at a reasonable cost. However, while new methodologies have been devised to handle and assemble these data (Ekblom and Galindo, 2011), methodological advances to convert these data into meaningful biological knowledge are still lagging behind. Indeed, one of the main challenges for the decade ahead will be to unravel the connection between genomic changes and the diversity of phenotypes seen in nature and to decipher the evolutionary forces responsible for this diversification. Such research efforts will lead to a more explanatory theory of evolution, with implications for all branches of the life sciences, from agriculture to ecology to medicine. Discovering these linkages is a difficult task that requires novel biologically-inspired computational methodologies that will identify candidate loci under selection and will correlate these loci to phenotypic differences.

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 1.1; pp. 1.1:1–1.1:21

A book completely handled by researchers.
No publisher has been paid.

In the post-genomic era, probabilistic models of molecular evolution are the powerhouse of data analyses. Such models form the backbone of various bioinformatics applications such as phylogenetic reconstruction (see Chapters 1.2 [Stamatakis and Kozlov 2020] and 1.4 [Lartillot 2020]), sequence alignment (see Chapter 2.2 [Ranwez and Chantret 2020]), molecular dating (see Chapter 5.1 [Pett and Heath 2020]), and functional sites prediction (Anisimova et al., 2013; Yang, 2014). These models are an attempt to represent the stochastic nature of evolution, built within a robust statistical framework of inference. Constructing evolutionary models enables us to mathematically describe various biological phenomena, estimate relevant parameters such as the strength of selection (see Chapter 4.5 [Lowe and Rodrigue 2020]) and rate of evolutionary events (see Chapter 4.4 [Bromham 2020]), and test different hypotheses, e.g., to determine which of several alternative evolutionary pathways is more likely. A main challenge in designing a model is to capture the key elements of the biological process at hand without over-parameterizing the model, which will render it inadequate. In molecular phylogenetics, Markov models, a specific class of stochastic models, are intensively used to analyze sequence data and to quantify the evolutionary dynamics of genes and genomes. In this chapter we introduce the theoretical foundations of these probabilistic models, starting with the simplest ones and progressing towards richer and more realistic models.

2 Discrete Time Markov chains

It is often helpful to start thinking of probabilistic evolutionary models in terms of simulations, i.e., to describe how one would mimic the evolutionary process using a computer. Specifically, we are interested in describing how sequences change through time and which parameters control this evolutionary process. For simplicity, we describe the evolution of only a single sequence site; assuming that all sites evolve independently, this process can be repeated N times to generate a sequence of length N. We start by drawing the identity of this position in the initial generation – the ancestor. Assuming that the probabilities of all nucleotides are equal, this is equivalent to rolling a dice with four possible outcomes (A, C, G, or T). Next, we would like to decide on the fate of this position in the next generation. It can either mutate or not. If it mutates, it can change to each one of the other nucleotides with equal probabilities. We can put this model into a matrix form:

$$M_{ij} = \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix} \tag{1}$$

This matrix represents a simple discrete-time Markov model with four states. The model is called discrete, because t, the time that corresponds to the number of generations, can only take discrete values $(0, 1, 2, \cdots)$. M_{ij} is the probability to start with nucleotide i and end in nucleotide j after one generation. Because each row includes all possible events, the sum of each row must equal one. There is one free parameter in this model, $0 \leq \epsilon \leq \frac{1}{3}$, which dictates the rate of evolution. If we run the simulations for many generations, a fewer mutations are expected for low values of ϵ compared to larger values. In fact, in each generation there are two possibilities: either a mutation has occurred, with probability 3ϵ or the position remained the same, with probability $1-3\epsilon$. Thus, the expectation of the number of changes for a single generation per position is 3ϵ.

Suppose now that we use this model to simulate 1,000 generations and that we are interested in calculating the probability that we end up with the nucleotide C in generation 1,000, given that we started with nucleotide A at generation zero. A naïve approach would be to simulate the evolution process for 1,000 generations according to the matrix M and record the final nucleotide (note that we always start the simulation with an A at generation zero). This represents one possible evolutionary outcome. In order to estimate the desired probability we repeat this process multiple times (say 1,000,000). The frequency of the outcome "ending with C" serves as a good estimation for the desired probability.

However, we can use Markov chain theory to analytically compute this probability. We first compute (rather than simulate) the probability of the event starting with A and after two generations ending with C. We denote this probability $P_{A,C}(2)$. The course of this event can be described as "A will change to some unknown character X after one generation" ($P_{A,X}(1)$) and then X will change to C in one generation ($P_{X,C}(1)$). By the law of total probability, we can express this probability by summing over all possible values of X:

$$P_{A,C}(2) = P_{A,A}(1)P_{A,C}(1) + P_{A,C}(1)P_{C,C}(1) + P_{A,G}(1)P_{G,C}(1) + P_{A,T}(1)P_{T,C}(1) \quad (2)$$

We note that we implicitly assume that the probability to change from A in generation 0 to X in generation 1, is the same as the probability to change from A in generation 1 to X in generation 2. More formally, we assumed that the Markov chain is time homogenous. In probabilistic terms, let $X(k)$ represent the nucleotide after k generations. Matrix M represents the transition probability from generation g to generation $g+1$ for every g and using this notation, we can express M as

$$M_{ij} = P(X(k+1) = j | X(k) = i) = P_{i,j}(1) \quad (3)$$

Note that this equality directly stems from the time homogeneity assumption. In matrix notations, we will hence consider M as $P(1)$, i.e., the Markov matrix representing the probabilities of changes after one generation. Note that we also implicitly assume that the transition probability from A in generation 1000 to C in generation 1001 is only dictated by M, and does not depend on the history of events that have occurred in generations 0 to 1000. This is an important property of the Markovian process, which is called memorylessness.

Returning to Eq. 2, one could notice that this computation is identical to the dot product of row A (first row) and column C (second column) in M. Let us denote by $P(2)$ the matrix of transition probabilities between each pair of nucleotides after two generations (and in general, $P(k)$ the matrix after k generations). The computation of each entry is performed in an identical way to Eq. 2:

$$P(2) = \begin{pmatrix} P_{AA}(2) & P_{AC}(2) & P_{AG}(2) & P_{AT}(2) \\ P_{CA}(2) & P_{CC}(2) & P_{CG}(2) & P_{CT}(2) \\ P_{GA}(2) & P_{GC}(2) & P_{GG}(2) & P_{GT}(2) \\ P_{TA}(2) & P_{TC}(2) & P_{TG}(2) & P_{TT}(2) \end{pmatrix} \quad (4)$$

From the above consideration, $P(2) = P(1)^2 = M^2$ and in general $P(k) = P(1)^k$ and $P(n+m) = P(n)P(m)$. The importance of the last equality will become clearer in the next section. Thus, in a discrete time process, by knowing the matrix M, the transition probabilities $P(k)$ can be derived for all possible k values. Hence, $M = P(1)$ is sometimes called the generator matrix for discrete-time Markov chains. Putting it all together, we obtain:

$$P_{ij}(k) = P(X(k) = j | X(0) = i) = \left[M^k\right]_{i,j} \quad (5)$$

3 Continuous Time Markov chains

A natural generalization of the above discrete time Markov chain is to replace the number of generations with a continuous parameter t that measures time. Indeed, such a process in which t can have any value in the interval $[0, \infty)$ is termed a continuous time Markov chain. Similar to discrete processes, $X(t)$ represents the state of a specific site at time t. Here, we would like to compute not only $P(1)$ and $P(2)$ but also, say, $P(2.335)$. To understand how such a matrix can be computed, we recall that for the discrete Markov process we had $P(t_1+t_2) = P(t_1)P(t_2) = P(t_2)P(t_1)$. This is also true for continuous time Markov chains, i.e., it is true for every two time points t_1 and t_2. From this, we obtain:

$$P(t_1+t_2) - P(t_1) = P(t_1)P(t_2) - P(t_1) = P(t_1)(P(t_2) - I) \tag{6}$$

where I is the identity matrix. Note that $P(0) = I$ since $P_{ii}(0) = P(X(0) = i|X(0) = i) = 1$ and similarly, $P_{ij}(0) = 0$ for all $i \neq j$. The above equations imply that for any $t_2 \neq 0$:

$$\frac{P(t_1+t_2) - P(t_1)}{t_2} = \frac{P(t_1)(P(t_2) - I)}{t_2} \tag{7}$$

We can then write t instead of t_1 and Δt instead of t_2. When Δt approaches zero, we obtain:

$$\lim_{\Delta t \to 0} \frac{P(t+\Delta t) - P(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t)(P(\Delta t) - P(0))}{\Delta t} = \lim_{\Delta t \to 0} P(t)\frac{P(\Delta t) - P(0)}{\Delta t} = P(t)P'(0) \tag{8}$$

where the last equality stems from the definition of the derivative at time zero. The left side of the above equation is the definition of the derivative at time t, and thus:

$$P'(t) = P(t)P'(0) \tag{9}$$

We note that in the equation above we made use of the concept of matrix derivative: if for a matrix $P(t)$ each element of a matrix is a function of t, then in $P'(t)$ each element is the derivative of the corresponding entry of $P(t)$. $P'(0)$ is a fixed matrix called the instantaneous rate matrix, and is often denoted as $Q \equiv P'(0)$.

As will be detailed below, given Q we can compute the transition probability matrix $P(t)$ for any time interval t, and thus Q is often termed the generator matrix of the continuous Markov process. The elements of each row in $P(t)$ sum to 1:

$$P_{i,1}(t) + P_{i,2}(t) + P_{i,3}(t) + \cdots + P_{i,n}(t) = 1 \tag{10}$$

where n is the number of possible states (four for nucleotides). Deriving both sides of this equation with respect to t, we obtain:

$$P'_{i,1}(t) + P'_{i,2}(t) + P'_{i,3}(t) + \cdots + P'_{i,n}(t) = 0 \tag{11}$$

Specifically, for $t = 0$, we obtain:

$$P'_{i,1}(0) + P'_{i,2}(0) + P'_{i,3}(0) + \cdots + P'_{i,n}(0) = 0 \tag{12}$$

Since $Q = P'(0)$ we obtain

$$Q_{i,1} + Q_{i,2} + Q_{i,3} + \cdots + Q_{i,n} = 0 \tag{13}$$

Thus, the elements in each row of Q sum to zero. For $i \neq j$, $P_{i,j}(t)$ is an increasing function of t: it is zero for $t = 0$ and non-negative for $t > 0$. Thus, each non-diagonal element, $Q_{i,j} = P'_{i,j}(0)$ is non-negative and represents the instantaneous transition rate from state i to state j, while each diagonal element, $Q_{i,i}$ is negative and equals to the negative sum of all other elements in that row. $Q_{i,i}$ represents the total instantaneous transition rate away from state i.

Eq. 9 above is in fact a differential equation in a matrix form:

$$\frac{dP(t)}{dt} = P(t)Q \tag{14}$$

The solution to this equation, subjected to the boundary condition $P(0) = I$ is

$$P(t) = e^{Qt} = I + Qt + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \cdots \tag{15}$$

We will note that this power series always converges. Computing matrix exponentials is a well-studied topic in numerical analysis and will not be discussed here. However, for certain types of simple models, a closed-form solution for $P(t)$ can be obtained directly. Below we derive the $P(t)$ matrix for the simplest nucleotide model, the Jukes and Cantor (JC) model.

4 The Jukes and Cantor model

The simplest continuous time Markov model for nucleotides assumes that the transition probabilities between each two different nucleotides is the same: $f(t)$. In a matrix form, $P(t)$ is:

$$P(t) = \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} \tag{16}$$

By the definition of Q for this model we obtain:

$$Q = P'(0) = \begin{pmatrix} -3f'(0) & f'(0) & f'(0) & f'(0) \\ f'(0) & -3f'(0) & f'(0) & f'(0) \\ f'(0) & f'(0) & -3f'(0) & f'(0) \\ f'(0) & f'(0) & f'(0) & -3f'(0) \end{pmatrix} \tag{17}$$

If we denote $f'(0)$ as α we obtain:

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \tag{18}$$

The stochastic process generated by this matrix (and assuming equal frequencies for all nucleotides) is called the Jukes and Cantor model (Jukes and Cantor, 1969). Applying Eq. 14 to the JC model, we obtain:

$$\frac{d}{dt}\begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} =$$

$$= \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} \cdot \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \quad (19)$$

If we focus only on the element in the first row and second column of the left side of this equation (and marking all the other elements as "*") and use the definition of matrix multiplication we obtain:

$$\begin{pmatrix} * & f'(t) & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} = \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \cdot \begin{pmatrix} * & \alpha & * & * \\ * & -3\alpha & * & * \\ * & \alpha & * & * \\ * & \alpha & * & * \end{pmatrix} \quad (20)$$

which yields:

$$f'(t) = (1-3f(t))\alpha + f(t)(-3\alpha) + f(t)\alpha + f(t)\alpha \quad (21)$$

Which, after simplification, results in the following regular differential equation:

$$\frac{df(t)}{dt} = \alpha(1-4f(t)) \quad (22)$$

This equation can informally be written as:

$$\frac{df(t)}{1-4f(t)} = \alpha 4dt \quad (23)$$

Integrating both sides yields:

$$\frac{ln(1-4f(t)}{-4}) = \alpha t + C \quad (24)$$

Isolating the term $f(t)$, we obtain:

$$f(t) = \frac{1}{4} - \frac{e^{-4\alpha t - 4C}}{4} \quad (25)$$

Recall that $f(t)$ is in fact $P_{i,j}(t)$ for $i \neq j$. Given that $f(0) = 0$ (because $P(0) = I$), we obtain that $C = 0$ and thus

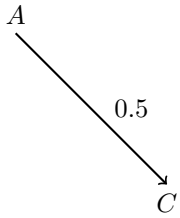
$$f(t) = \frac{1}{4} - \frac{e^{-4\alpha t}}{4} \quad (26)$$

Similarly, since $P_{i,i}(t)$ is $1 - 3f(t)$ we obtain:

$$P_{i,j}(t) = \begin{cases} \frac{1}{4} - \frac{e^{-4\alpha t}}{4} & \text{if } i \neq j, \\ \frac{1}{4} + \frac{3e^{-4\alpha t}}{4} & \text{if } i = j \end{cases} \quad (27)$$

5 Likelihood functions and matrix normalization

In the JC model, given the α and t parameters, we can simulate an initial sequence and let it evolve, *in silico*. We can also write explicit probabilistic expressions for evolutionary scenarios. We start with a nucleotide sequence composed of a single position, in which the character A is observed. The probability of starting with A according to the JC model is $P(A) = 0.25$ since the JC model assumes equal probabilities for all nucleotides. Of note, this implies that out of all simulations starting with a sequence of a single position, on average 1/4 would start with the character A. Now, consider the case of a position in which the ancestral character is A and the descendant character, after time $t = 0.5$, is C. We graphically denote this scenario by:



The probability of simulating such a scenario under the model is $P(A) \cdot P_{A,C}(0.5)$. If we consider the two sequences as our data, D, and the model as M, we can denote this probability by $P(D|M)$, i.e., the probability of the data given the model and its parameters. This is called the likelihood of the model. Consider now the more complicated case in which the sequence ACGGT evolved to ACGAC, again along time $t = 0.5$. The likelihood of the model is the product of the probabilities of each position in the data:

$$(P(A)\cdot P_{A,A}(0.5))\cdot(P(C)\cdot P_{C,C}(0.5))\cdot(P(G)\cdot P_{G,G}(0.5))\cdot(P(G)\cdot P_{G,A}(0.5))\cdot(P(T)\cdot P_{T,C}(0.5)) \tag{28}$$

Note that the product of probabilities reflects the assumption that given the model parameters, the tree and its branches, positions evolve independently of one another. This assumption is a clear oversimplification of the biological reality, but it is included in the vast majority of models employed in molecular evolution since it allows rapid computation of the likelihood function. Under the JC model, the likelihood depends on both t and α (each $P_{i,j}(t)$ term depends on both t and α). We can thus term it $L(t, \alpha)$.

When examining natural biological sequences, we do not know the values of the parameters that gave rise to them. As common in many statistical applications, inferring these values often involves finding the parameter values that maximize the likelihood function. These values are called MLEs: maximum-likelihood estimates. Note however that there are infinitely many (t, α) pairs that maximize the likelihood function. This is because in all these expressions, t and α always appear together as a single term, αt. Thus, we can always multiply α by a constant value, and divide t by this same constant value and the likelihood would remain unchanged. This is a well-known phenomenon in statistics and such models are termed non-identifiable, i.e., there is no one-to-one mapping between the likelihood function and a set of model parameter values. Note that this non-identifiability is not restricted to the JC model; from Eq. 15, we see that $P(t)$ depends on Qt. Thus, we can always multiple Q by a factor and divide t by that same factor, and the $P(t)$ function (and consequently the likelihood function) would remain unchanged. However, in spite the infinite number of MLE (t, α) pairs, we can uniquely infer their product αt. For the above example, and using the JC model, $\alpha t = 0.19$ maximizes the likelihood function.

We will now see that the αt product is also related to the average number of substitutions when sequences evolve for t units of time.

For continuous time evolutionary Markov processes, d, the expected number of substitutions along a branch of length t, can be computed by the following expression:

$$d = -t \sum_i P(i) \cdot Q_{i,i} \tag{29}$$

We do not show the derivation of this formula here. For the above JC model, this yields $d = 3\alpha t$. This suggests that we can choose any one of the pairs (t, α) that maximizes the likelihood function, compute d based on this pair, and obtain the expected number of substitutions along the branch associated with the MLEs. Alternatively, we can set $\alpha = 1/3$, and that would imply that $d = t$. The resulting JC model is:

$$P_{i,j}(t) = \begin{cases} \frac{1}{4} - \frac{e^{-4t}}{4} & \text{if } i \neq j, \\ \frac{1}{4} + \frac{3e^{-4t}}{4} & \text{if } i = j \end{cases} \tag{30}$$

This can be generalized beyond the JC model as according to Eq. 29, d would be equal to t if

$$\sum_i P(i) \cdot Q_{i,i} = -1 \tag{31}$$

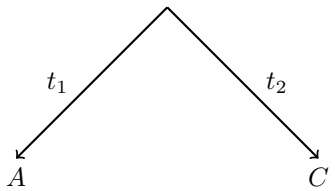
We can always multiply the Q matrix by a fixed factor so that Eq. 31 holds. This is called matrix normalization and, following this procedure, estimates of t would in fact correspond to the expected number of substitutions along a branch of length t, rather than to time.

6 Reversibility, homogeneity and stationarity

Time reversibility of a Markov process imposes the following equation:

$$P(i) \cdot P_{i,j}(t) = P(j) \cdot P_{j,i}(t) \tag{32}$$

This suggests that the probability of starting with character A and ending up with character C along a branch of length t is the same as the probability of starting with character C and ending up with character A (along the same branch). More generally, it suggests that when we have two sequences and we assume that the first is the ancestor and the second is the descendent, the probability of obtaining the two sequences would have been the same if we had rather assumed that the first is the descendent and the second the ancestor. Even more generally, we can choose the "oldest" (ancestral root) point anywhere along the branch t. This point divides the branch t into two segments t_1 and $t_2 = (t - t_1)$, each of which connects the sequences to that new root. Reversibility enforces that the likelihood function becomes invariant to the position of the root, i.e., it would remain the same regardless of our choice of t_1. We show it here for the very simple case of two single character sequences:



The division of the branch between the sequences defines the simplest bifurcating tree. The likelihood computation along the tree is given by summing over all possible states x of the root:

$$\sum_x P(x) \cdot P_{x,A}(t_1) \cdot P_{x,C}(t_2) \tag{33}$$

Using the reversibility condition for the first two terms we obtain:

$$\sum_x P(A) \cdot P_{A,x}(t_1) \cdot P_{x,C}(t_2) \tag{34}$$

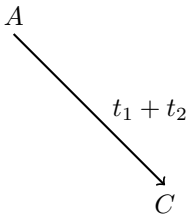
We can take the first term outside the summation:

$$P(A) \sum_x \cdot P_{A,x}(t_1) \cdot P_{x,C}(t_2) \tag{35}$$

Using the Markovian property, we can simplify the equation to:

$$P(A) P_{A,C}(t_1 + t_2) \tag{36}$$

We thus see that the likelihood depends on the total length of two branches, suggesting that one can move the root position as long as the total length remains fixed. This means that we can set one branch to zero and the other one to $t_1 + t_2$ and the likelihood would remain unchanged, i.e., one sequence can be considered ancestral to the other one:



For an unrooted tree with three taxa, it is possible to place the root at any point along each one of the three branches. The same argument that shows that all root positions results in the same likelihood for two sequences holds also for this case, and in fact, the reversibility condition makes all rooted trees that are derived from the same unrooted tree equally probable. This suggests that when assuming a reversible model, it is impossible to infer the location of the tree root. However, reversibility allows very efficient algorithms for inferring branch lengths and thus, for searching for the tree with the highest likelihood, by using the so called pulley principle (Felsenstein, 1981).

If we look at a discrete (or continuous) time Markov chain, we are often interested in the probabilities of being at a specific state after the chain has been running for a long time. In some cases, we find that as t approaches infinity, $P(t)$ converges to a specific form in which all rows are identical, such that $P(\infty) = \pi \cdot I$, and π is a row vector of size n, in which π_i is the probability of reaching character i after infinite amount of time. Moreover, the vector of probabilities π is unchanged by the transition matrix $P(1)$, i.e., $\pi \cdot P(1) = \pi$. Such cases are called stationary chains and π is a unique vector of stationary probabilities. Once the chain reaches stationarity, it remains in stationarity. The JC model, for example, has stationary probabilities of exactly 0.25 for each character.

Homogeneity means that a single Q matrix characterizes the entire evolutionary process. In contrast, branch heterogeneous models assume that different branches of a tree have different Q matrices.

Currently, the vast majority of phylogenetic studies assume models that are reversible, stationary, and homogenous. However, it is clear that such models are oversimplification of reality in many cases. For example, bacterial genomes vary substantially in their GC composition across homologous regions, in clear violation of these assumptions (Galtier and Lobry, 1997).

7 Basic models of nucleotide substitutions

As stated above, in the JC model, which is represented by the simplest nucleotide transition probabilities (see Eq. 27), it is assumed that the instantaneous substitution rates between all pairs of different nucleotides are identical, which implies that the stationary frequencies equal 0.25. There is, however, ample empirical evidence that substitution probabilities vary among nucleotide pairs. A plethora of extensions to this basic modeling scheme have been developed over the last 50 years, each implying different hypotheses regarding the pattern of nucleotide evolution. The Kimura two parameters model (K2P, Kimura, 1980) alleviates the often unrealistic assumption that transitions (a substitution between two pairs of purines or between two pairs of pyrimidines) and transversions (a substitutions between a purine and a pyrimidine or vice versa) are equiprobable. This has resulted is the following Q matrix, where α is the rate of transitions and β is the rate of transversions:

$$Q[K2P] = \begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{pmatrix} \tag{37}$$

As for the JC mode, matrix normalization introduces a constraint on the matrix, i.e., given a specific α value, one can compute the β value. In fact, it is possible to rewrite the K2P matrix, following normalization, using a single parameter: the transition-transversion rate ratio. Which of the infinite many possible transition-transversion rate ratios should be used? The one that maximizes the likelihood function. We call this parameter "free", because it is estimated using likelihood based on the analyzed data. While other types of parameterizations are possible, each capturing a different biological aspect, the richest model possible, under the assumption that substitution rates are symmetrical (i.e,, $Q_{i,j} = Q_{j,i}$ for all i and j) is captured by a matrix with six parameters (five free parameters) and is denoted as the SYM model (Zharkikh, 1994):

$$Q[SYM] = \begin{pmatrix} -\alpha-\beta-\gamma & \beta & \alpha & \gamma \\ \beta & -\beta-\delta-\epsilon & \delta & \epsilon \\ \alpha & \delta & -\alpha-\delta-\eta & \eta \\ \gamma & \epsilon & \eta & -\gamma-\epsilon-\eta \end{pmatrix} \tag{38}$$

Because the JC, K2P, and SYM matrices are symmetrical ($Q_{i,j} = Q_{j,i}$ for all i and j) the stationary frequencies of all nucleotides are equal to 0.25. As stated above, biological dataset are often characterized with biased nucleotide frequencies. An important extension to such models relies on non-symmetrical matrices that can results in any desired stationary nucleotide frequencies, denoted by π. To this end, three additional free parameters are added to the model: π_A, π_C, and π_G, representing the frequency of A, C, and G (the frequency of T, π_T is constrained such that the sum of all probabilities is 1). Incorporating these parameters

into the JC, K2P, and SYM models resulted in an expanded set of models, termed F81 (Felsenstein, 1981), HKY (Hasegawa et al., 1985), and the General Time Reversible model (GTR, Tavaré, 1986), respectively, and in general, models that incorporate this possibility are usually denoted by "+F". To retain the time reversibility property, such models take the following general form

$$Q[GTR] = \begin{pmatrix} -\dots & \beta\pi_C & \alpha\pi_G & \gamma\pi_T \\ \beta\pi_A & -\dots & \delta\pi_G & \epsilon\pi_T \\ \alpha\pi_A & \delta\pi_C & -\dots & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_C & \eta\pi_G & -\dots \end{pmatrix} \tag{39}$$

Note that the model is time reversible because it satisfies $\pi_i Q_{i,j} = \pi_j Q_{j,i}$ for all i and j.

8 Basic codon models

The models detailed above all assumed that the only possible characters are the four DNA nucleotides, i.e., we use an alphabet of size four. When protein-coding nucleotide sequences are analyzed, the transition rates are highly influenced by the effect they exert on the encoded protein. In this case, transition rates between codons I and J may depend on the type of nucleotide substitution, the type of amino-acid replacement, and by codon usage (i.e., the frequency of the target codon). The use of codon models was originally developed around the same time by Goldman and Yang (1994) and Muse and Gaut (1994), see also Chapter 4.5 (Lowe and Rodrigue, 2020). The model is generally represented by a 61×61 codon rate matrix, which describes the instantaneous substitution rate from codon $I = i_1 i_2 i_3$ to codon $J = j_1 j_2 j_3$. This matrix assigns different rates to nucleotide substitutions at the three codon sites according to their type and effect on the coded amino acid:

$$Q_{I,J} = \begin{cases} \alpha_{i_k j_k}\pi_J & I \text{ and } J \text{ differ by one synonymous substitution at codon site } k \\ r(A_I, A_J)\alpha_{i_k j_k}\pi_J & I \text{ and } J \text{ differ by one nonsynonymous substitution at codon site } k \\ 0 & I \text{ and } J \text{ differ by more than one nucleotide} \end{cases} \tag{40}$$

Here, π_J is the frequency of codon J and $\alpha_{i_k j_k}$ is the substitution rate between nucleotide i_k to nucleotide j_k, which can be parameterized based on any time reversible nucleotide substitution model, such as GTR or HKY. $r(A_I, A_J)$ is the replacement rate between the amino acids A_I and A_J encoded by codon I and J, respectively. Note that to maintain the reversibility property, $r(A_I, A_J)$ should be equal to $r(A_J, A_I)$. In the model originally suggested Goldman and Yang (1994), $r(A_I, A_J)$ was dependent on two components: (1) the physiochemical distance between the amino acids A_I and A_J, as measured by the matrix provided by Grantham (1974), and (2) on a free parameter that accounts for the intensity of selection acting on the encoded protein. In the MG model of Muse and Gaut (1994), $r(A_J, A_I)$ was represented simply by the selection-intensity parameter ω, which specifies the nonsynonymous to synonymous substitution rate ratio:

$$Q_{I,J} = \begin{cases} \alpha_{i_k j_k}\pi_J & I \text{ and } J \text{ differ by one synonymous substitution at codon site } k \\ \omega\alpha_{i_k j_k}\pi_J & I \text{ and } J \text{ differ by one nonsynonymous substitution at codon site } k \\ 0 & I \text{ and } J \text{ differ by more than one nucleotide} \end{cases} \tag{41}$$

9 Basic amino-acid models

The analysis of sequences at the amino acid level requires an instantaneous rate matrix of dimension 20 over 20. If the substitution rate between each pair of amino acids is considered a free parameter, 190 parameters need to be estimated from the analyzed data (189 after matrix normalization), which is both computationally demanding and requires large amounts of data for accurate inference. Thus, for protein sequences, a pre-computed matrix is usually used for which the parameters were inferred based on a very large protein dataset. For example, Dayhoff et al. (1978) curated an atlas of all the available protein sequences at that time, and estimated the substitution rates based on a maximum-parsimony-like procedure. When more data became available and when inference procedures improved, updated matrices were inferred, e.g., the JTT, the WAG, and the LG matrices (Jones et al., 1992; Le and Gascuel, 2008; Whelan and Goldman, 2001). In addition, matrices for specific datasets were also introduced: mtREV for mitochondrially encoded proteins (Adachi and Hasegawa, 1996), cpREV for chloroplast encoded proteins (Adachi et al., 2000), and similarly, matrices for different secondary structures or surface accessibility (Koshi and Goldstein, 1995). These matrices are considered empirical because they are based on averaging substitution rates across many datasets. This is in contrast to the above mechanistic models, in which the model parameters were chosen to reflect certain assumptions regarding the substitution pattern and are estimated for each dataset. Of note, these matrices can be decomposed into two components:

$$Q = S \cdot \Pi \tag{42}$$

where S is a symmetrical matrix describing the amino-acid exchangeability component and the diagonal matrix Π represents the amino-acid stationary frequencies. These 20 amino acid frequencies can be estimated from each analyzed data (the "+F" option), and thus, amino-acid models are often a mix between a component estimated based on a very large sequence compendium (the "S" component) and a dataset-specific component (the stationary amino acid frequencies).

10 Among-site-rate-variation

The above models assume the exact same stochastic process at each sequence site. This should not be taken to mean that all positions would experience the same number of substitutions. Due to the stochastic nature of the evolutionary process, given a phylogenetic tree and a specific model, by chance alone, some sequence sites experience more substitutions than others. These differences in the number of substitutions per site follow a Poisson distribution (Yang, 1996). However, it is now well established that the distribution of the number of substitutions per site in real biological data is substantially different from what can be expected by chance alone, i.e., different from a scenario in which all positions evolve exactly under the same stochastic model. For example, for a given dataset, there is often an excess of invariant positions (i.e, they are fully conserved and experience no substitutions along the phylogeny), compared to data simulated assuming a constant model across sites. Branch lengths reflect the number of substitutions averaged over all sites. Thus, if the average number of substitutions per site across the entire tree is high, sites that experience no substitutions are expected to be extremely rare. However, in protein sequences, for example, there are many more invariant sites than the expected number. These sites are usually those that are critical for maintaining the function or structure of the protein,

so that purifying selection removes most of the mutations that appear at these positions. Models that aim to capture the observed variability of substitution rates among different sites are called among-site-rate-variation (ASRV) models. To understand how they work, we first introduce the concept of site-specific evolutionary rate. Consider a site that evolves along a branch of length t, say under the JC model. Because the branch lengths are indicative of the average number of substitutions per site, a site that evolves along a branch of length t is expected to accumulate half the number of substitutions compared to a site that evolves along a branch of length $2t$. Consider the case, in which in a specific position the ancestor character is A and the descent character is C. The likelihood of these data given a branch length of $2t$ is $P_{A,C}(2t)$. An equivalent way to think of this case is of a character evolving along a branch of length t, with a site-specific rate of 2. In general, the likelihood of a scenario in which sequence $A_1, A_2, \ldots, A_N$ evolves into the sequence $B_1, B_2, \ldots, B_N$ along a branch of length t and with site specific rates $r_1, r_2, \ldots, r_N$ is:

$$\prod_{i=1}^{N} P_{A_i B_i}(r_i \cdot t) \tag{43}$$

In general, the site-specific rates are unknown. One possibility would be to assign each site with its own rate parameter. This, however, would result in inevitably large number of parameters that have to be estimated from the data, and would generally result in inferior inferences (Mayrose et al., 2004). Instead, we can assume that these rates are taken from a limited set of values $r_1, r_2, \ldots, r_k$, with corresponding probabilities $w_1, w_2, \ldots, w_k$. Thus, these rates are sampled from a discrete distribution. We can then compute the likelihood while averaging over all possible rates:

$$\prod_{i=1}^{N} \sum_{j=1}^{k} P_{A_i B_i}(r_j \cdot t) \cdot w_j \tag{44}$$

The question then becomes how to choose these rates and their probabilities. Tamura and Nei (1993) were the first to suggest that, for pairwise sequences, rates are gamma distributed. Yang (1993) showed how to compute the likelihood of a tree assuming the rates are sampled from a continuous gamma distribution. Later, Yang (1994) showed how assuming the rates are sampled from a discretized version of the gamma distribution can speed up computations. By far, the discrete gamma distribution is the most widely used ASRV distribution. The gamma distribution with parameters α and β has mean α/β and variance α/β^2. Usually, the unit 1 gamma distribution is used such that $\alpha = \beta$ and the mean rate over all sites is 1. Models incorporating this possibility thus include a single additional parameter, α, and are usually denoted by "+G". In the discretized version of these models, it is common to choose the representative rates, such that each has equal probability $1/k$ (Yang, 1994). While alternative approaches for the discretization of the gamma distribution, based on Laguerre quadrature for example, were suggested (Felsenstein, 2001; Mayrose et al., 2005), these are rarely used. Additionally, several alternatives to the gamma distribution were proposed. First, an additional category, specific for invariant sites was proposed, generating the G+I model (Gu et al., 1995). Second, Kosakovsky Pond and Frost (2005) suggested to discretize the gamma distribution into categories based on quantiles estimated using a discretized beta distribution. Third, Mayrose et al. (2005) suggested that a mixture of several discrete gamma distributions better captures ASRV compared to using a single gamma distribution. Finally, Yang (1995) suggested a free parameter distribution, in which both the rates and their probabilities are parameters which are estimated using maximum likelihood. While

this distribution is highly flexible and can well approximate data-specific characteristics, it is also very parameter rich. Introducing ASRV into probabilistic evolutionary models was shown to increase the accuracy of tree inference and branch lengths as well as many downstream applications that rely on evolutionary models such as molecular dating (Yang, 1996) and quantifying site specific conservation scores (Mayrose et al., 2004). For a review regarding modeling ASRV, see Yang (1996) and Pupko and Mayrose (2010).

11 Mixture models

The above models assume that a single Q matrix represents the evolutionary dynamics across all sites. However, the selective forces and possibly the mutation processes may vary among sites, and in this sense, such models may be an oversimplification of the evolutionary process. It is possible to alleviate this restriction by assuming that there are several possible instantaneous rate matrices Q_1, Q_2, ..., Q_N and that each site is associated with one of these matrices. In case we do not know the matrix that is associated with each site, we compute the likelihood by averaging over all possible matrix assignments, weighted by their probabilities:

$$P(D|M) = \sum_{j=1}^{k} P(D|M_j)P(M_j) \tag{45}$$

where M_j is the model defined by using the matrix Q_j. This is very similar to likelihood computations that incorporate site-specific rates , as detailed in the previous section, where here we average over the various Q matrices rather than averaging over the possible rate categories. Both of these cases are considered mixture models (Zhang and Huang, 2015). Mixture models are widely used to describe codon evolution. While in the MG codon model described in Eq. 41 a single ω parameter is assumed for all sites, it is clear that the type and intensity of this selection coefficient vary, with some sites experiencing purifying selection ($\omega < 1$) while some sites evolve neutrally ($\omega = 1$) or possibly, under positive selection ($\omega > 1$). Mixture models allow modeling this variability in selection intensity directly. A set of possible ω values is assumed, which results in a set of Q matrices. The likelihood is then computed using the above formula for mixture models. Various mixture models were suggested for codon models, in which the ω values vary according to either a free, a gamma or a beta distribution (Yang et al., 2000). Using these models it is possible to test for the presence of positive selection, by contrasting the likelihoods of a null mixture model that allows only Q matrices with $\omega \leq 1$ and an alternative mixture model that also includes a Q matrix with $\omega > 1$. Such models are also often used to infer the posterior estimates of ω for each site, thus revealing sites that evolve under specific selection regimes (Cannarozzi and Schneider, 2012).

12 Gene family models

The above models consist of continuous time Markov chains of nucleotide, amino acid and codon sequences. However, Markov models were also developed to describe the evolutionary dynamics of gene families. The simplest form of such models allows the analysis of phyletic patterns, in which the dataset is a matrix representation of gene family presence and absence across a set of genomes, in which each row represents a genome, each column represents a gene family, and the i, j entry in the matrix is 1 if a member of gene family j is present in genome i and 0 otherwise. The states of such a model are binary $\{0, 1\}$, and thus, a two by

two Q matrix is used to model the evolution of these characters. In such a matrix, $\lambda = Q_{0,1}$ is the gain rate and $\mu = Q_{1,0}$ is the loss rate:

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \tag{46}$$

We note that a gain event may reflect either de-novo emergence of a gene family or cases of gains by obtaining a copy of a gene via horizontal gene transfer. We also note that a '1' to '0' transition reflects the loss of all copies of a given gene family. Using this rate matrix, it is possible to derive explicit formulae for $P_{ij}(t)$ (Ross, 1996). Imposing reversibility on such a matrix, i.e., $\pi_0\lambda = \pi_1\mu$ results in the condition $\mu = \lambda\pi_0/\pi_1$ and we obtain an alternative representation:

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \lambda\pi_0/\pi_1 & -\lambda\pi_0/\pi_1 \end{bmatrix} = \frac{\lambda}{\pi_1}\begin{bmatrix} -\pi_1 & \pi_1 \\ \pi_0 & -\pi_0 \end{bmatrix} = r_Q \begin{bmatrix} -\pi_1 & \pi_1 \\ \pi_0 & -\pi_0 \end{bmatrix} \tag{47}$$

Here, r_Q is a scaling parameter. Modifying its value would only change the total number of transitions along the phylogenetic tree but not the relative number of gains and losses. If we impose the constraint that branch lengths reflect average number of substitutions along the tree, we should enforce $\sum_i \pi_i Q_{i,i} = -1$, which for this case yields $r_Q = 1/(2\pi_0\pi_1)$ and since $\pi_0 + \pi_1 = 1$, this model has only a single free parameter: π_0. The above model was used and extended in various studies. Hao and Golding (2006) assumed that the gain and loss rates are equal, thus ensuring that the size of the genome does not vary much during evolution. Cohen et al. (2008) extended this model by allowing unequal gain and loss rates as well as introducing rate variation among gene families using a discrete gamma distribution. In such a model, the rate varies but the gain and loss rate ratio is assumed to be identical among all gene families. Subsequent development alleviated this assumption of fixed gain and to loss ratio by introducing mixture models (Cohen and Pupko, 2010; Spencer and Sangaralingam, 2009). One interesting aspect of such models is the need to correct for unobservable data when analyzing empirical data. Consider a case where the ancestral gene family was lost in all its descendants. In the phyletic pattern, this corresponds to a column of zero, reflecting a gene family that is absent from all present-day genomes. However, in empirical data, columns of zero are never observed because phyletic patterns are constructed by homology searches among present-day genomes. Felsenstein (1992) already noted and suggested a solution for this problem when analyzing restriction site presence-absence data. Specifically, the likelihood of the observed data for a specific gene family (D) is in fact conditioned on not being a column of zeros. Let C_0 and C_+ be the events "a column of zeros" and "not a column of zeros", respectively. We obtain:

$$P(D|C_+) = \frac{P(D\&C_+)}{P(C_+)} = \frac{P(D)}{1 - P(C_0)} \tag{48}$$

Thus, the desired probability, $P(D)$ is the product of the conditional probability and the probability of a column not made entirely of zeros. The latter can be easily estimated by computing the probability of the complementary event–a column of zeros. In the above continuous time Markov models, the state "1" represents one or more copies. Instead, it is possible to extend such models to explicitly account for the number of copies in each gene family. In theory, the number of states is infinite but in practice a pre-defined upper bound, M, is used to transfer the state space into a finite state Markov chain, such that the last state includes all values equal or above that number. In this case, the data are coded over the alphabet 0, 1, 2,..., M. The rate matrix in this case is a variant of a birth death model

(Ross, 1996), where the birth rate, λ, and the death rate, δ, are the rates of a single gene gain or loss:

$$Q_{i,j} = \begin{cases} \lambda & \text{j=i+1} \\ \delta & \text{j=i-1} \\ 0 & \text{otherwise} \end{cases} \tag{49}$$

Here too, several extensions have been proposed that allowed for: (i) de novo emergence of a gene family, as specified by the birth-death-innovation model (Csurös, 2010; Karev et al., 2003; Librado et al., 2012; Spencer et al., 2006), (ii) the possibility of whole genome duplication (Rabier et al., 2014), (iii) dependence of the gain and loss rates on the number of gene families (Spencer et al., 2006), and (iv) accounting for differential sequencing coverage and quality of annotations across genomes (Han et al., 2013).

13 Indel models

While probabilistic substitution models are now routinely used when reconstructing phylogenetic trees or searching for positive selection, for the inference of multiple sequence alignments, ad-hoc methods are commonly employed. This difference stems from the fact that modeling insertion and deletion (indel) events is more challenging compared to modeling substitutions, mainly because incorporating indel events within the likelihood function violates the assumption that different sites evolve independently, which is required for efficient likelihood computations. In a breakthrough paper, Thorne et al. (1991) developed the first probabilistic model that includes indel events (the TKF91 model). Unlike substitution models, in which the number of states in the Q matrix is 4, 20 or 61 (for nucleotides, amino acids, or codons), in TKF91, the sequences need to be considered as whole and cannot be seen as "independent columns". Thus, the number of states is exponential in the length of the sequence. This large number of possible states makes direct exponentiation of an explicit Q matrix impossible. To reduce the complexity of the model, the TKF91 model assumes that $Q_{i,j} = 0$ if the length difference between sequences i and j is greater than one. This assumption implies that longer length differences are the outcome of several indel events of length one. While this assumption is clearly unrealistic from what we know of indel mutations, it was made in order to make computations with this model feasible. The introduction of Bayesian integration techniques in TKF91 as well as advanced dynamic programming algorithms further enhanced the ability to compute with such a probabilistic indel model. The TKF91 was extended by the same group of researchers to allow longer indels (TKF92, Thorne et al., 1992). However, to overcome computational challenges, it was assumed that overlapping indels never occur through evolution, which is again biologically unrealistic. A full long-indel model was developed by Miklos et al. (2004). While this model allows indels of any size and overlaps, it is extremely computationally intensive and cannot be applied for more than a handful of sequences. Recently, Levy Karin et al. (2018) introduced significant accelerations to this long-indel model and demonstrated that using such an indel model results in more accurate pairwise alignments compared to widely-used alignment programs, such as MAFFT (Katoh and Standley, 2013), which do not rely on an explicit probabilistic model. Statistical alignment algorithms aim at the joint inference of trees and alignments using probabilistic models (Steel and Hein, 2001). Current tools for statistical alignments such as BaliPhy (Redelings and Suchard, 2005) rely on hidden Markov Models (HMMs), which are also probabilistic-based indel models. However, implicitly, in HMM-based models there is a distinct Markov model for each tree branch. In contrast,

the indel models presented above describe a single process over the whole tree, in which a single set of model parameters is shared among all tree branches. While indel-based models currently lag behind substitution-only models, their development holds the promise to dramatically change various molecular-evolution applications, such as sequence alignment algorithms, the characterization of indel evolutionary dynamics in various genes and lineages (Chen et al., 2009; Levy Karin et al., 2017a; Lunter, 2007), algorithms for simulating sequences (Cartwright, 2005; Fletcher and Yang, 2009) as well as downstream analyses such as tree inference and molecular dating.

14 More sophisticated models

With the increased availability of sequence data and the increased computational resources, the development of more sophisticated inference procedures of sequence evolution has undergone accelerated evolution in itself. A very partial list of influential directions in model development includes:

1. models that account for variation of the process among tree branches. Such model include, for example, codon models that allow for positive selection only on a subset of tree branches (see Yang et al. (2002); Chapter 4.5 [Lowe and Rodrigue 2020]) and models that allow the rate of evolution to change along the tree, e.g., the covarion models (Galtier, 2001);
2. more sophisticated mixture models, which allow averaging over a set of empirical amino-acid matrices (Quang et al., 2008) or sampling amino acid matrices using a Bayesian approach (see also Lartillot and Philippe (2004); Chapter 1.4 [Lartillot 2020]);
3. models that integrate protein structure information with sequence evolution (Choi et al., 2007);
4. models that integrate trait information with sequence evolution (Lartillot and Poujol, 2011; Levy Karin et al., 2017b);
5. models in which the substitution rate continuously evolves (Lartillot and Poujol, 2011);
6. models that allow different partitions of the datasets to evolve under different sets of parameters (Nylander et al., 2004; Lanfear et al., 2016).

References

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of molecular evolution*, 42(4):459–68.

Adachi, J., Waddell, P. J., Martin, W., and Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of molecular evolution*, 50(4):348–58.

Anisimova, M., Liberles, D. A., Philippe, H., Provan, J., Pupko, T., and von Haeseler, A. (2013). State-of the art methodologies dictate new standards for phylogenetic analysis. *BMC evolutionary biology*, 13(1):161.

Bromham, L. (2020). Substitution rate analysis and molecular evolution. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.4, pages 4.4:1–4.4:21. No commercial publisher | Authors open access book.

Cannarozzi, G. M. and Schneider, A. (2012). *Codon Evolution: Mechanisms and Models*. Oxford University Press.

Cartwright, R. A. (2005). DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics (Oxford, England)*, 21 Suppl 3(Suppl 3):iii31–8.

1.1:18 REFERENCES

Chen, J. Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., and Tian, D. (2009). Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Molecular Biology and Evolution*, 26(7):1523–1531.

Choi, S. C., Hobolth, A., Robinson, D. M., Kishino, H., and Thorne, J. L. (2007). Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular Biology and Evolution*, 24(8):1769–1782.

Cohen, O. and Pupko, T. (2010). Inference and characterization of horizontally transferred gene families using stochastic mapping. *Molecular biology and evolution*, 27(3):703–13.

Cohen, O., Rubinstein, N. D., Stern, A., Gophna, U., and Pupko, T. (2008). A likelihood framework to analyse phyletic patterns. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1512):3903–11.

Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics (Oxford, England)*, 26(15):1910–2.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In Dayhoff, editor, *Atlas of protein sequence and structure*, volume 5 suppleme, pages 345–352. National Biomedical Research Foundation, Washington.

Ekblom, R. and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol*, 17(6):368–76.

Felsenstein, J. (1992). Phylogenies From Restriction Sites: A Maximum-Likelihood Approach. *Evolution*, 46(1):159–173.

Felsenstein, J. (2001). Taking variation of evolutionary rates between sites into account in inferring phylogenies. *Journal of Molecular Evolution*, 53(4-5):447–455.

Fletcher, W. and Yang, Z. (2009). INDELible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8):1879–1888.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, 18(5):866–873.

Galtier, N. and Lobry, J. (1997). Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes. *Journal of Molecular Evolution*, 44(6):632–636.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.)*, 185(4154):862–4.

Gu, X., Fu, Y. X., and Li, W. H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution*, 12(4):546–57.

Han, M. V., Thomas, G. W. C., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular biology and evolution*, 30(8):1987–97.

Hao, W. and Golding, G. B. (2006). The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research*, 16(5):636–643.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS*, 8(3):275–82.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *Academy Press*, pages p. 21–132.

Karev, G. P. G., Wolf, Y. Y. I., and Koonin, E. E. V. (2003). Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics*, 19(15):1889–1900.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–20.

Kosakovsky Pond, S. L. L. and Frost, S. D. D. (2005). A simple hierarchical approach to modeling distributions of substitution rates. *Mol Biol Evol*, 22(2):223–234.

Koshi, J. M. and Goldstein, R. A. (1995). Context-dependent optimal substitution matrices. *Protein engineering*, 8(7):641–5.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2016). PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Molecular biology and evolution*, page msw260.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Lartillot, N. and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol*, 28(1):729–744.

Le, S. Q. and Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, 25(7):1307–1320.

Levy Karin, E., Ashkenazy, H., Hein, J., and Pupko, T. (2018). A simulation-based approach to statistical alignment. *Systematic Biology*.

Levy Karin, E., Shkedy, D., Ashkenazy, H., Cartwright, R. A., and Pupko, T. (2017a). Inferring rates and length-distributions of indels using approximate Bayesian computation. *Genome Biology and Evolution*, 9(5):1280–1294.

Levy Karin, E., Wicke, S., Pupko, T., and Mayrose, I. (2017b). An Integrated Model of Phenotypic Trait Changes and Site-Specific Sequence Evolution. *Systematic Biology*, 66(6).

Librado, P., Vieira, F. G., and Rozas, J. (2012). BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics (Oxford, England)*, 28(2):279–81.

Lowe, C. and Rodrigue, N. (2020). Detecting adaptation from multi-species protein-coding dna sequence alignments alignments. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.5, pages 4.5:1–4.5:18. No commercial publisher | Authors open access book.

Lunter, G. (2007). Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, 23(13).

Mayrose, I., Friedman, N., and Pupko, T. (2005). A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, 21 Suppl 2:ii151–8.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol*, 21(9):1781–1791.

Miklos, I., Lunter, G. A., and Holmes, I. (2004). A "Long Indel" Model for Evolutionary Sequence Alignment. *Molecular Biology and Evolution*, 21(3):529–540.

Muse, S. and Gaut, B. (1994). A likelihood approach for comparing synonymous and non-synonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724.

Nylander, J., Ronquist, F., Huelsenbeck, J., and Nieves-Aldrey, J. (2004). Bayesian Phylogenetic Analysis of Combined Data. *Systematic Biology*, 53(1):47–67.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Pupko, T. and Mayrose, I. (2010). Probabilistic Methods and Rate Heterogeneity. In *Elements of Computational Systems Biology*. Wiley Online Library.

Quang, L. S., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323.

Rabier, C.-E., Ta, T., and Ané, C. (2014). Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Molecular biology and evolution*, 31(3):750–62.

Ranwez, V. and Chantret, N. (2020). Strengths and limits of multiple sequence alignment and filtering methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.2, pages 2.2:1–2.2:36. No commercial publisher | Authors open access book.

Redelings, B. D. and Suchard, M. A. (2005). Joint bayesian estimation of alignment and phylogeny. *Systematic Biology*, 54(3):401–418.

Ross, S. M. (1996). *Stochastic Processes*. John Wiley & Sons, New York, NY, 2nd edition.

Spencer, M. and Sangaralingam, A. (2009). A Phylogenetic Mixture Model for Gene Family Loss in Parasitic Bacteria. *Molecular Biology and Evolution*, 26(8):1901–1908.

Spencer, M., Susko, E., and Roger, A. J. (2006). Modelling prokaryote gene content. *Evolutionary bioinformatics online*, 2:157–78.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Steel, M. and Hein, J. (2001). Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree. *Applied Mathematics Letters*, 14(6):679–684.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–26.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences. Volume 17*, pages 57–86.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of molecular evolution*, 33(2):114–24.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. *Journal of molecular evolution*, 34(1):3–16.

Whelan, S. and Goldman, N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*, 18(5):691–699.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular biology and evolution*, 10(6):1396–401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol*, 39(3):306–14.

Yang, Z. (1995). A space-time process model for the evolution of dna sequences. *Genetics*, 139(2):993–1005.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in ecology & evolution*, 11(9):367–72.

Yang, Z. (2014). *Molecular evolution: a statistical approach*. Oxford University Press.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449.

Yang, Z., Nielsent, R., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19(6):908–917.

Zhang, H. and Huang, Y. (2015). Finite Mixture Models and Their Applications: A Review. *Austin Biom and Biostat. Austin Biom and Biostat*, 2(2):1013–1.

Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of molecular evolution*, 39(3):315–29.

Chapter 1.2 Efficient Maximum Likelihood Tree Building Methods

Alexandros Stamatakis
Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies,
Karlsruhe Institute of Technology, Institute for Theoretical Informatics
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
Alexandros.Stamatakis@h-its.org
https://orcid.org/0000-0003-0353-0691

Alexey M. Kozlov
Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
Alexey.Kozlov@h-its.org
http://orcid.org/0000-0001-7394-2718

Abstract

The number of possible unrooted binary trees (phylogenies) increases super-exponentially with the number of taxa. To find the Maximum Likelihood (ML) tree one has to enumerate and evaluate all these trees. As we will see, this is computationally not feasible. Therefore, one predominantly deploys *ad hoc* tree search methods that strive to find a "good" ML tree in the hope that it will be close, either with respect to the likelihood score or the topological structure, to the globally optimal ML tree. In this chapter we provide an overview over the most popular and efficient ML tree search techniques.

How to cite: Alexandros Stamatakis and Alexey M. Kozlov (2020). Efficient Maximum Likelihood Tree Building Methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 1.2, pp. 1.2:1–1.2:18. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Funding This work was funded by the Klaus Tschira Foundation.

1 Introduction

The number of possible phylogenetic trees grows super-exponentially with the number of taxa. In many cases such a combinatorial explosion means that the optimization problem, that is, finding the ML tree, is what is called *NP-hard* in computer science.

In simple words, NP-hardness means that there does not exist any known algorithm for solving the problem requiring polynomial runtime as a function of the input size. In our case, the input size is the number of taxa and number of sites of the input, that is, the Multiple Sequence Alignment (MSA) of the taxa for which we desire to infer a tree. Throughout this chapter we will assume that the MSA is given.

Using the machinery of theoretical computer science, it has been formally proved that finding the optimal tree for character-based tree scoring criteria such as parsimony (Day, 1987) and ML (Roch, 2006) is NP-hard. In other words, we will need to calculate the ML score of every single possible tree to find the ML tree.

With the computer power available today, this might, at first glance not appear to be problematic. Assume, however, that we want to infer a ML tree on a MSA with 100 taxa which is, by current standards, only a medium-sized dataset with respect to the number of

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 1.2; pp. 1.2:1–1.2:18


taxa. For 100 taxa there exist roughly 1.7×10^{182} distinct unrooted phylogenies as calculated by our `TreeCounter` tool that we have found helpful for teaching purposes (available at https://github.com/stamatak/TreeCounter).

The exact number of possible unrooted binary trees for 100 taxa is:

170045880929340962283784787050354160735772501841042490022783586836362580888628332485131901009696411611113290250954694628264213300391989811682923929339908722247494604289531707763671875

Further, for the sake of simplicity, assume that calculating the ML score on one tree takes 1 second. To find the ML tree, one has to score all 1.7×10^{182} phylogenies. This requires an overall runtime of roughly 5×10^{174} years. In turn, this corresponds to only about the 3.8×10^{164}-fold age of our universe. It is important to note that, using powerful parallel supercomputers does not help to reduce this comparatively long waiting time as supercomputers reduce running times linearly (w.r.t., the number of processors they have) at best, but unfortunately, not super-exponentially. Assuming that we could use a rather large supercomputer with 10^{12} processors[1], we would still have to wait for the 3.8×10^{152}-fold age of the universe for the ML tree.

As we presumably do not want to wait for this long, we need to devise heuristic search strategies that navigate through this enormous space of phylogenies in an "intelligent" way and return a tree with a "good" score. As mentioned before, most of the heuristics currently used are *ad hoc* strategies. In other words, they do not offer any theoretical guarantees of how close or far away (regarding the ML score) the tree they return is from the globally optimal tree. This is also the reason why the sloppy terminology that we often observe in empirical evolutionary biology papers is potentially misleading. Papers often refer to "the ML tree". However, this is simply the best phylogeny found by the completely *ad hoc* heuristic search strategy.

Given the prolegomena, one might wonder what the computational complexity of Bayesian Inference of phylogenies might be, since it essentially also relies on repeated likelihood evaluations on trees (see Chapter 1.4 [Lartillot 2020a]). One would assume that their complexity must also somehow be affected by the vastness of tree space. If we simply look at Bayes' equation:

$$P(T|D) = \frac{P(T) \times P(D|T)}{P(D)} \tag{1}$$

where T is the tree, D the data, $P(T)$ is the prior probability and $P(D|T)$ the standard phylogenetic likelihood (as used for ML inference) score of the tree. The computational problem is hidden in $P(D)$ that represents the marginal probability of the data which cannot be computed exactly but needs to be approximated. An exact evaluation of this term would, again, as for ML, require calculating the likelihood scores of all possible topologies and, in the Bayesian setting, also for all possible remaining parameter values (e.g., branch lengths, rates of nucleotide substitution etc.). So calculating $P(D)$ exactly would require even longer waiting times than for ML above. As a consequence, one would rather not opt to calculate the posterior probability $P(T|D)$ exactly. As a work-around, the posterior probability is approximated via Markov-Chain Monte-Carlo (MCMC) methods. Unfortunately, these MCMC chains are only guaranteed to converge to the true posterior distribution if they are

[1] Evidently, no such supercomputer exists yet.

run for infinity. In practice, the lack of MCMC convergence can be assessed using so-called convergence analysis tools (e.g. Nylander et al., 2008). It can not be emphasized frequently enough that, these methods can not be used to demonstrate that the chains have converged. Based on the mathematical foundations of MCMC the chains will only converge to the true posterior probability distribution if executed eternally. Convergence analysis tools can therefore only detect lack of convergence. If there is no lack of convergence, then what they do indicate is that the MCMC process has reached an area of *apparent* convergence.

In the following sections we will introduce ML tree search algorithms in a top-down fashion. Initially, we discuss the commonalities of ML search strategies in Section 2. Subsequently we discuss the most common mechanisms for changing tree topologies in Section 2.2 and also cover some important implementation details (Section 2.3). In Section 3 we discuss how search strategies use these tree change moves to find "good" trees. We also discuss why we think that divide-and-conquer strategies for ML-based tree inference have not been successful to date (Section 3.1) and how terraces in tree space affect tree searches (Section 3.2).

In Section 4 we discuss the computation of support values using the standard phylogenetic bootstrap procedure and take a critical look at some recently proposed methods for obtaining approximate support values. We conclude in Section 5 with some notes of caution and recommendations on how to best infer a phylogeny from our point of view with a focus on selecting the best search strategy. For the sake of clarity, we deliberately omitted other important topics such as model selection or the selection of the most appropriate partitioning scheme.

2 Top-Level View of Search Algorithms

Initially, we will discuss the basic components that form part of almost every ML-based tree search algorithm. Most search strategies comprise the following two steps:

1. Construct an initial comprehensive tree that contains all taxa of the input MSA and compute its ML score
2. Start changing this initial comprehensive topology to find a tree with a better ML score

The most widely-used tools for ML-based inference (RAxML (Stamatakis, 2014), IQ-Tree (Nguyen et al., 2015), PHYML (Guindon et al., 2010), and GARLI (Zwickl, 2006)) implement this strategy. However, researchers have also experimented with divide and conquer approaches where the comprehensive tree is assembled by puzzling together independently optimized smaller subtrees. We discuss these approaches in more detail in Section 3.1.

2.1 Constructing Comprehensive Trees

There are at least three solutions to constructing a comprehensive starting tree: random topology, neighbor-joining (NJ) or its variants such as BioNJ (Gascuel, 1997), and parsimony. As predominantly done in Bayesian phylogenetic inference (since by definition a randomized stochastic sampling should start at a random point), one can simply construct a complete random starting tree. Alternatively, one might opt to infer a somewhat reasonable tree using a simpler tree building method such as neighbor-joining (NJ). A NJ starting tree typically has a better likelihood score than a complete random tree. On the other hand, just using the NJ tree, might drive the subsequent tree search into a local optimum. Because of the immense vastness of tree space that might exhibit a plethora of local optima, it is thus desirable to implement mechanisms (usually relying on some sort of randomization) that allow for navigating out of local optima (see Figure 8 for an example of a local optimum)

in some way. This can either be achieved by generating a distinct set of starting trees or via a randomization step in the tree search procedure. For now, we will focus on generating distinct starting trees. If our tool deploys complete random starting trees, obtaining a set of n distinct starting trees is straight-forward as we simply need to generate n comprehensive trees at random.

If we want to obtain a set of distinct starting trees with a better (i.e., non-random) initial likelihood score than a random tree, we can deploy the so-called randomized stepwise addition algorithm described below, using a simple criterion of our choice. Simple, in this context means, cheap-to-compute with respect to the computational cost of likelihood calculations. The parsimony or least-squares criteria are good examples of such simple criteria. Such reasonable yet simple criteria will generate starting trees that have a "good" initial likelihood score. Therefore, the subsequent tree search on the comprehensive tree (see Section 2.2) is likely to converge faster, hence reducing inference times. See Figure 1 for an example.

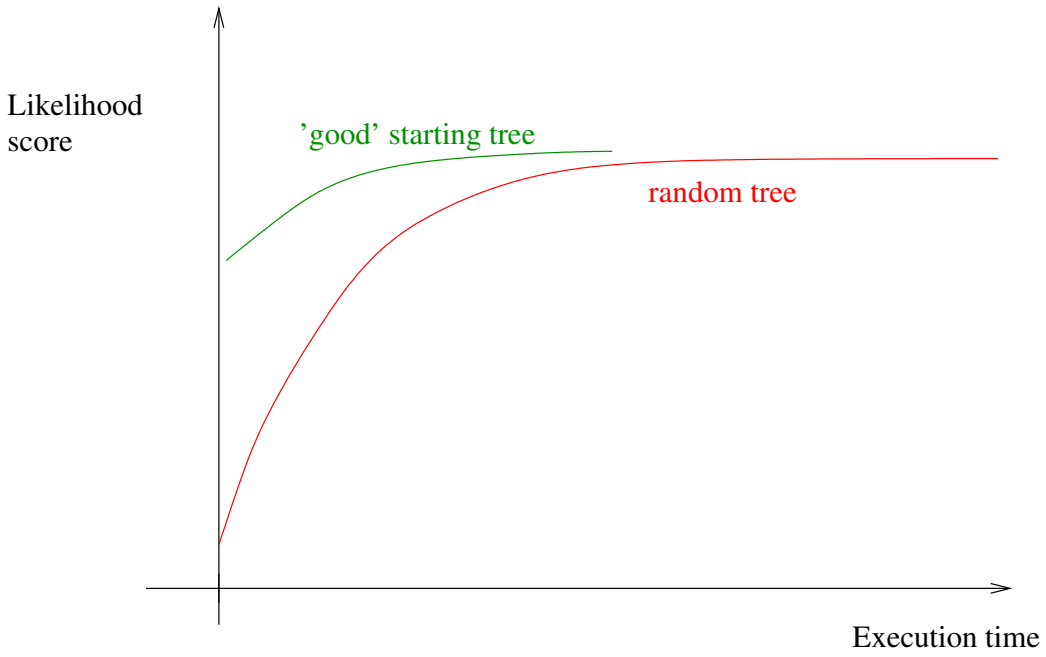


Figure 1 Schematic ML run time differences when initiating tree searches on random versus "good" comprehensive trees.

In the following, we will briefly outline the randomized stepwise addition order algorithm given a MSA with n taxa. The algorithm is also outlined in Figure 2.

1. Chose three taxa t_1, t_2, t_3 at random and use them to construct the only possible unrooted strictly binary three taxon tree
2. Chose the next taxon t_i to insert at random from the list of the remaining $n-3$ taxa
3. Insert t_i into every branch of the already constructed tree that has $i-1$ taxa. For each insertion of t_i into a branch calculate and store the score using, for instance, parsimony. Then, remove t_i from the current branch again.
4. Once, we have computed insertion scores for all branches, finally insert t_i into the branch that yielded the best insertion score.
5. Continue adding taxon after taxon to the tree as above until no taxa are left to insert.

In general, applying this procedure several times using distinct randomized taxon addition orders (e.g., inserting in this order t_1, t_2, t_3, t_4, t_5 versus inserting in the following order t_3, t_1, t_5, t_2, t_4) we will obtain a set of topologically distinct comprehensive initial trees. However, if the signal in the data is very strong (e.g., large concatenated supermatrices as in Misof et al. (2014) it might well be, and we have observed this while analyzing empirical datasets, that several distinct addition orders do yield the same tree. Therefore, one should first check, for instance, by computing the Robinson-Foulds distances (Robinson and Foulds, 1981) between all pairs of starting trees (e.g., using an appropriate script or `-rfdist` command of RAxML-NG), how many distinct trees the inferred starting tree set does contain, prior to launching computationally intensive ML searches on these trees.

One might wonder, why we typically do not use likelihood as a criterion for this randomized stepwise addition procedure. This is simply again due to the computational cost of ML. Note that, evaluating likelihoods on trees takes between 85%-95% of overall runtime in all likelihood-based tree inference tools (this also holds for Bayesian inference). In addition, using likelihood in this step does not yield substantially better trees than using parsimony (Morrison, 2007). Moreover, one will typically observe larger likelihood improvements by applying topological changes to the comprehensive tree (see Section 2.2 below). Thus, using parsimony or NJ/BioNJ represents a classic engineering trade-off between the quality (likelihood score) of the initial tree and the time required to construct it. For the sake of completeness, it is worth mentioning that some older tools, namely fastDNAml (Olsen et al., 1994) and PAUP* (Swofford, 2001), implemented a randomized stepwise addition procedure using Maximum Likelihood as an option.

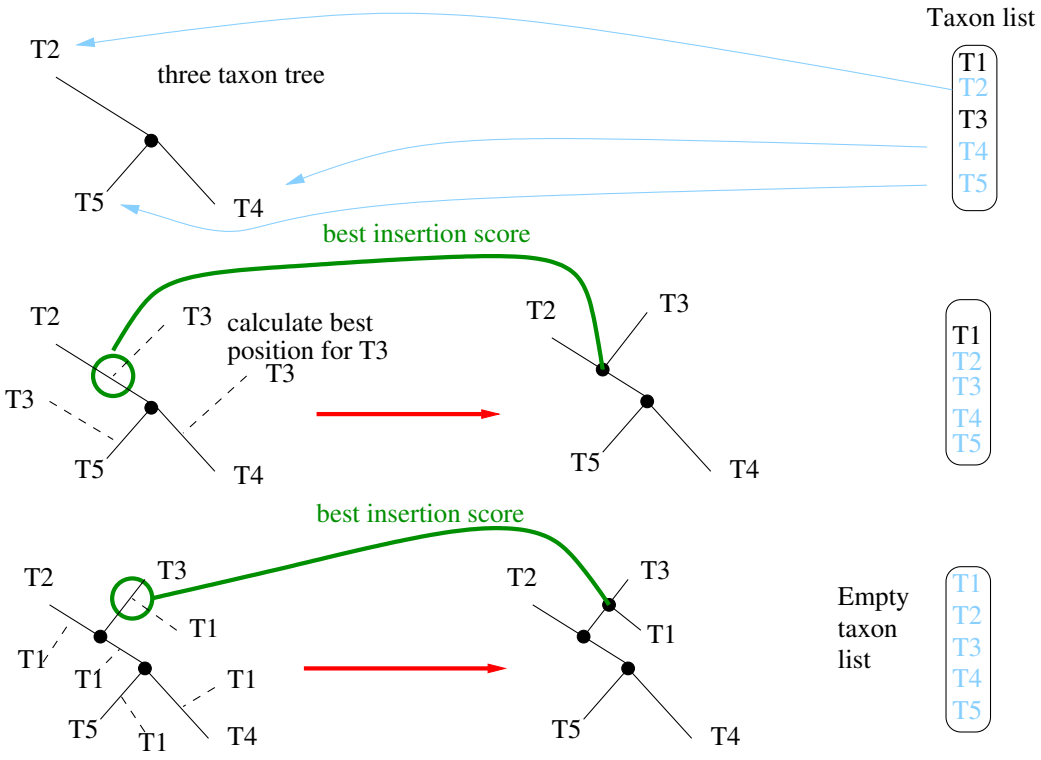


Figure 2 Example of randomized stepwise addition algorithm for constructing an initial tree comprising 5 taxa.

To conclude, using parsimony-based randomized stepwise addition order starting trees

(as in RAxML or IQ-Tree) might potentially bias the search towards specific parts of the tree space and particular local optima. Therefore, search strategies relying on comprehensive starting trees and subsequent greedy hill climbing can benefit from using different starting tree types (e.g., *both* random *and* parsimony) to more thoroughly explore tree space. In general, this needs to be assessed on a case by case basis, depending on the data at hand. The way we imagine the search space is visualized in Figure 3.

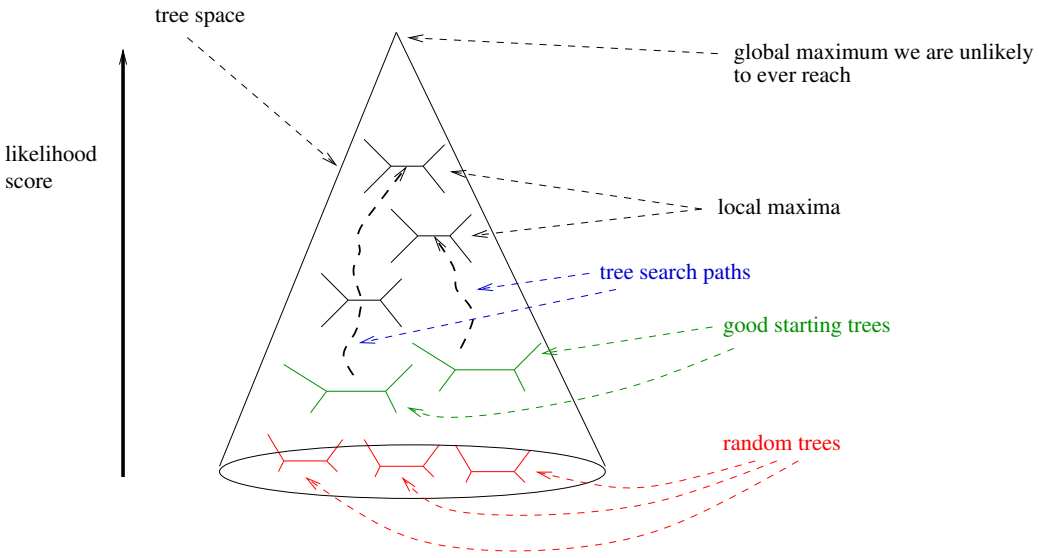


Figure 3 Our way of imagining tree search space, including random starting trees, "good" starting trees, and tree search paths that take us closer to the desired global maximum, that is *the* ML tree.

As already mentioned, most Bayesian inference programs typically start from a random tree. However, some implementations (MrBayes (Ronquist et al., 2012), ExaBayes (Aberer et al., 2014)) do offer the option to also initiate the MCMC procedure on a randomized stepwise addition order parsimony tree.

2.2 Changing Topologies - Searches on Comprehensive Trees

Now that we know how to compute a comprehensive starting tree, we can consider the basic techniques for changing that tree in order to further improve the likelihood.

We will first only consider the widely-used standard tree moves that allow us to change the tree to varying degrees. In other words, we will consider bold (change the tree topology substantially) versus conservative (slightly change the tree) topological alteration mechanisms. We will discuss how to use these standard mechanisms to build a tree search strategy in Section 3 including several examples. When discussing moves, we always refer to a given tree as the current tree. This can, for instance, be the best tree we have found so far and that is stored in memory. Via a tree move we then attempt to construct a new tree by changing the current tree in the hope that this new tree will have a better likelihood than the current tree.

The three most widely used fundamental tree moves (also referred to as topological alteration mechanisms) are the following:

1. NNI: Nearest Neighbour Interchange (see Figure 4)
2. SPR: Subtree Pruning and Re-grafting (see Figure 5)

3. TBR: Tree Bisection and Reconnection (see Figure 6)

The most simple as well as most conservative move is the NNI move. To apply a NNI move we first need to select (how we do this selection is a matter of designing the search strategy) an inner branch of the tree that defines a blue (B), a yellow (Y), a red (R), and a green (G) subtree. In our example in Figure 4 subtrees B and R are located on one side of the branch and Y and G on the other side. To generate alternative trees with NNI we can now flip our colored subtrees over the inner branch. Thereby, from the tree containing this inner branch, we can construct two alternative, not substantially different (in terms of topological distance to the initial tree; for bolder moves see SPR and TBR explained below), tree topologies and evaluate their likelihood scores.

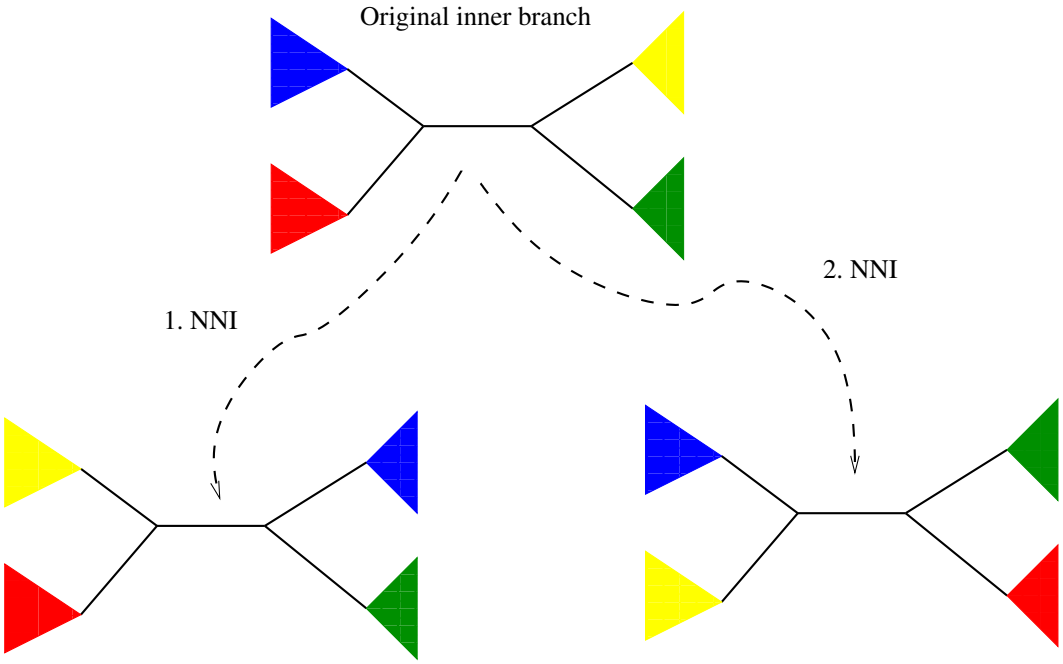


Figure 4 Outline of the two possible NNI moves as executed on an inner branch of a phylogenetic tree defining red, blue, green, and yellow subtrees.

A more bold move than the NNI is the SPR move. To carry out a SPR move, we first select the root of a subtree in the comprehensive tree. Again, how and in which order we select such a root depends on the actual tree search strategy. In RAxML, for instance, we conduct a depth first traversal of the tree and apply SPR moves to *all* subtrees we encounter. Once we have selected a subtree root, we initially prune the subtree (called *candidate subtree*) by removing its subtree root from the branch to which it is attached to. We call the original branch in the current tree where the subtree was pruned the *pruning branch*. Then we can start inserting, calculating the likelihood, and removing again our candidate subtree into the neighboring branches of the pruning branch. We can define the size of this neighborhood by the so-called rearrangement radius (see Figure 5). The rearrangement radius allows us to determine up to how many nodes away from the pruning branch we desire to re-insert our candidate subtree. If we set the rearrangement radius to 1, we will only execute conservative SPR moves, whereas if we set the rearrangement radius to the total number of inner nodes in the tree (remember that this is $n - 2$ for an unrooted binary tree with n taxa), we perform bold moves and explore a larger portion of the colossal tree space. How to set this

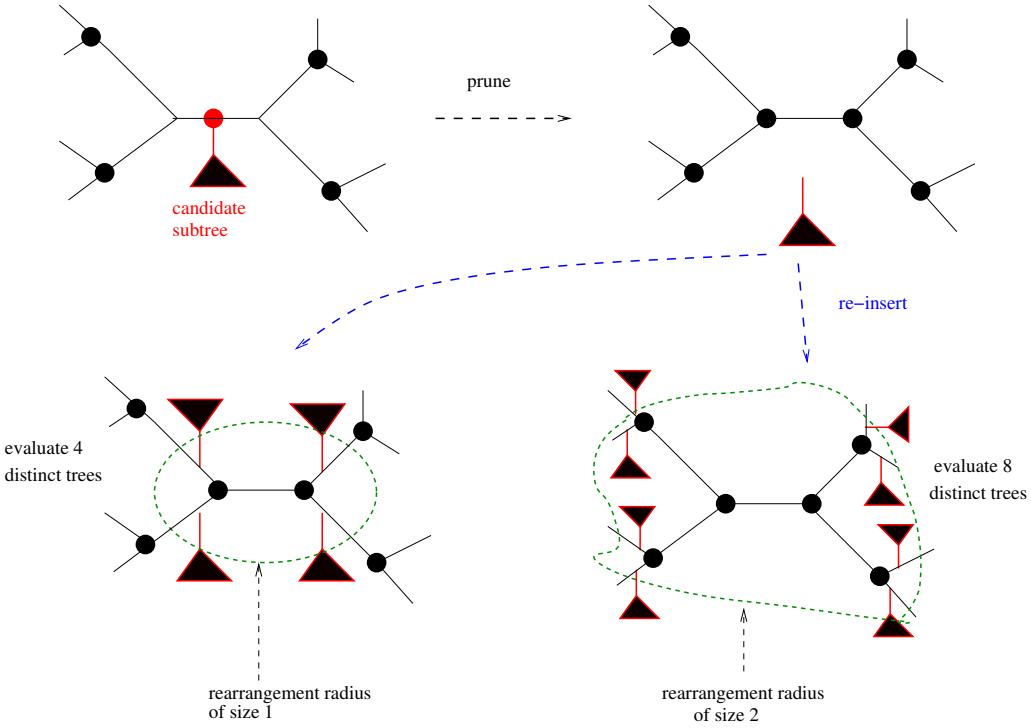


Figure 5 Outline of the Subtree Pruning and Re-Grafting Procedure. First, a candidate subtree is selected. Then, it is pruned from the current tree. Subsequently, one can re-insert it at all branches that are one node away from the pruning branch (rearrangement radius of 1), all branches that are two nodes away from the pruning branch (rearrangement radius of 2), etc.

rearrangement radius, how to adapt it to the MSA at hand, and how to potentially change it over the course of a tree search is again subject to search strategy development. In RAxML, for instance, the default behavior is to automatically determine a "good" rearrangement radius using the following strategy (implementation details omitted): choose the smallest SPR radius that yields the largest likelihood improvement on the starting tree. In most other tools, the rearrangement radius is set to a fixed default value, but can be changed by the user via respective command line flags.

The most drastic tree move is the TBR move (see Figure 6). A TBR move is drastic because it can induce large topological changes on the tree. As a consequence, a TBR move can also either substantially increase or decrease the likelihood of a tree. To execute a TBR move one initially selects a branch to *bisect* the current tree to obtain two unconnected subtrees t_1 and t_2. Then, one generates alternative tree topologies by reconnecting the two subtrees. This is accomplished by connecting all pairs of branches in the two unconnected subtrees via a new branch. Thereby, one obtains $n \times m$ alternative tree topologies, where n is the number of branches in t_1 and m the number of branches in t_2. As for SPR moves, one can limit the range of these reconnection operations to a neighborhood around the respective positions from which the original branch was removed.

2.3 Implementation Details

The three fundamental tree moves appear to be relatively straight-forward. To maximize computational efficiency, that is, to minimize the amount of phylogenetic likelihood calcu-

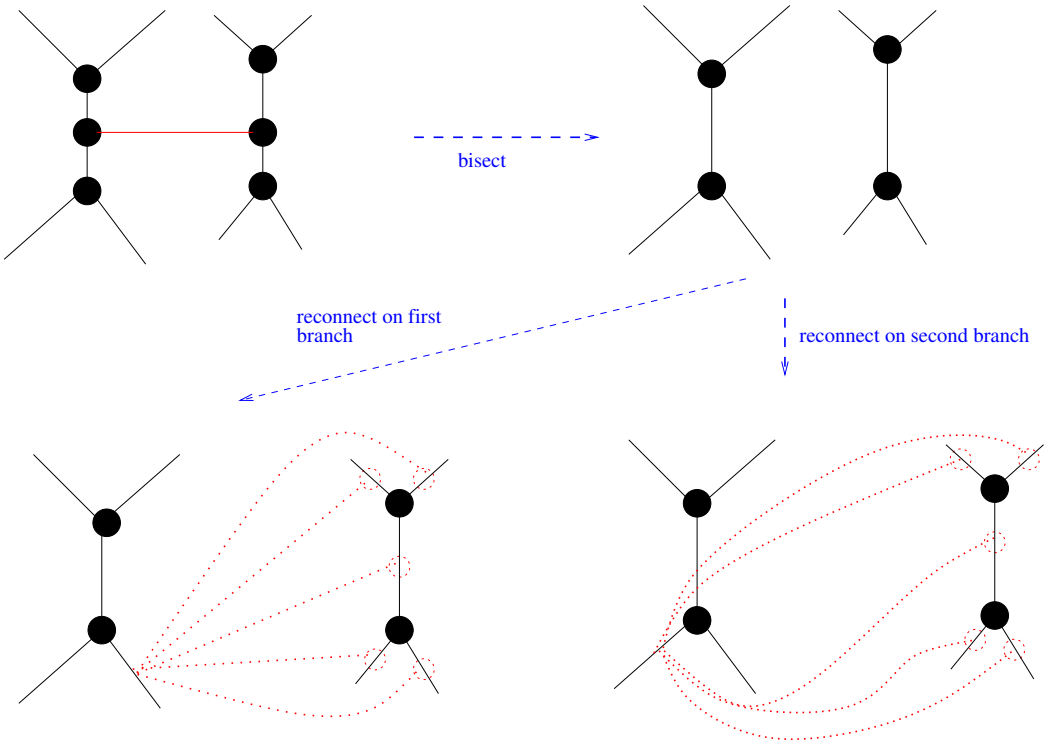


Figure 6 Outline of the tree bisection and reconnection move. The tree is initially bisected by removing the red branch in the top left corner of the figure to obtain two disjoint subtrees t_1 and t_2. Then, we can start reconnecting them by visiting each branch of the left subtree and connecting it with all branches of the right subtree as shown for two out of the five branches of the left subtree in the bottom of the figure.

lations, a plethora of shortcuts and heuristics are applied in practice. In principle, after each tree move, for instance, after applying a NNI move, one would need to re-optimize *all* branch length values and *all* remaining model parameters (e.g. GTR rates, α shape parameter of the Γ distribution to model rate heterogeneity) to obtain *the* ML score of the tree generated via the move. As such global optimization operations on phylogenies are highly compute-intensive because they require repeatedly traversing and re-computing the likelihood on the entire tree, all common ML based tree inference programs use shortcuts. In other words, they only compute an approximate likelihood score for a tree generated via a tree move and not *the* ML score.

Typically there are three types of shortcuts that are used by ML software developers.

2.3.1 Avoiding model parameter optimization

To circumvent the high computational cost of re-estimating the ML model parameters (GTR rates, α parameter) with the exception of the branch lengths, one relies on the following empirical observation: As long as the tree we are applying our moves to is reasonable (i.e., non-random) the model parameter estimates will not change substantially. Thus, it is sufficient to only re-estimate them periodically after having applied a relatively large number of tree moves and potential changes to the tree topology.

2.3.2 Avoiding global branch length optimization

To avoid re-estimating all branch lengths of the new tree after a move, program developers rely on the following intuition: The branch lengths that are in the neighborhood where the tree was changed should be affected most by the move. Hence, only those supposedly most affected branch lengths are typically re-estimated. Consider, for instance, a NNI move. Here, one assumes that the 5 branches shown in black in Figure 4 are affected most by the NNI move. Thus, one would only re-estimate these 5 branch lengths after applying a NNI. One could even decide to only re-estimate the center branch. What works best is a matter of numerous trial-and-error experiments on benchmark datasets during program development. For SPR moves one can apply the so-called lazy SPR move technique that was introduced in RAxML. Here, only the three branch lengths adjacent to the subtree insertion position are re-estimated (see Figure 7). In GARLI there is a more elaborate method for conducting lazy SPR moves. GARLI tries to dynamically determine the number of branches that need to be re-estimated after a move. In other words, it optimizes branch lengths at an increasing distance from the subtree insertion position until they do not change significantly any more.

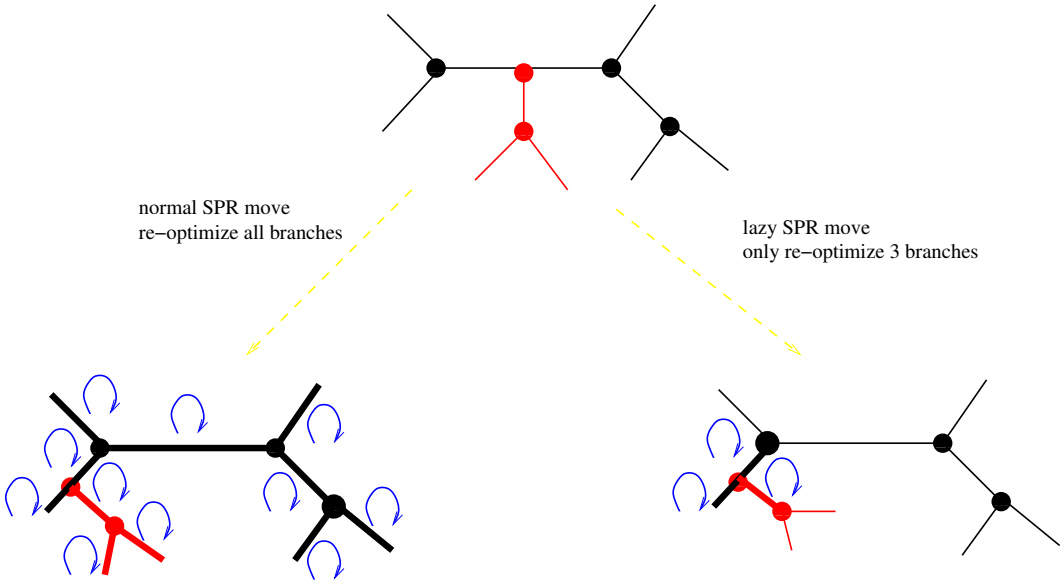


Figure 7 Outline of a standard versus a lazy SPR move on the subtree with the red branches. In the left hand bottom corner we conduct a standard SPR move, that is, we re-optimize *all* branch lengths of the tree (indicated by thick lines). In the right hand bottom corner we conduct a lazy SPR move, that is, we only re-optimize the three branches (shown by thick lines) that are adjacent to the insertion position of the subtree.

2.3.3 Locality of tree move applications

To avoid jumping back and forth between different distant regions of the tree while applying topological moves, most ML search algorithms apply moves systematically to the tree in a pre-defined order (e.g., a depth first traversal of the tree). This improves computational efficiency as the moves (and hence the costly updates of Conditional Probability Vectors) are always taking place in only one region of the tree, while the rest of the tree remains unaltered. Then, one moves on to a neighboring region (e.g., a neighboring subtree which one tries to rearrange with SPR moves).

3 Search Strategies

Given our set of fundamental tree moves, we can now design tree search strategies. Most popular search algorithms repeatedly apply one or several of these moves to the current tree until there is no move that yields a tree with a better ML score. Once, no move can be applied that further improves the ML score, the algorithm converges and returns the best-scoring ML tree it was able to find.

A simple search strategy as implemented in the original version of PHYML (Guindon and Gascuel, 2003)

1. Build a NJ starting tree.
2. Repeatedly apply NNI moves to all inner branches of the tree.
3. Terminate if for none of the inner branches there is a NNI move that can further improve the ML score of the tree.

The key pitfall of such a simple search strategy is that it is highly likely to get stuck early in a local optimum as outlined in Figure 8.

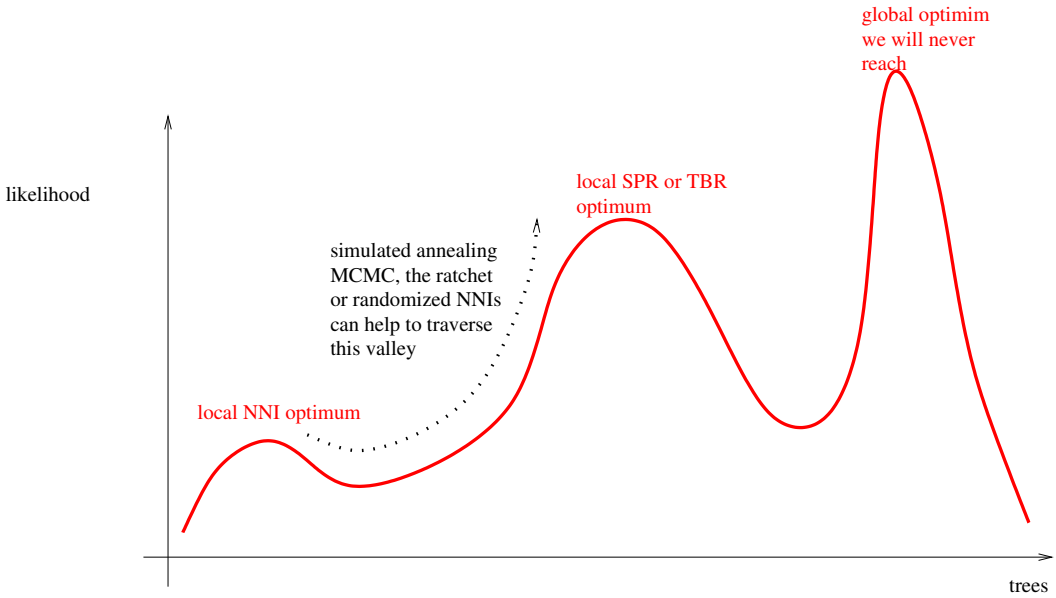


Figure 8 Example of how ML searches can get stuck in local optima. The first local optimum on the right could be a local NNI-optimum, that is, we are not able to navigate out of this optimum by applying NNI moves. One could move from this optimum to the better local optimum in the middle of the graph via SPR or TBR moves or by applying a stochastic search (e.g., simulated annealing or MCMC in the Bayesian setting. To the right we see the global optimum which we are unlikely to ever reach.

One way to alleviate this is to use more radical moves such as SPR or TBR that allow to more easily move away from such a local optimum again. RAxML and GARLI mainly rely on lazy SPR moves, while PHYML version 3 also introduced SPR moves (Guindon et al., 2010).

In addition, one can conduct multiple tree searches on several distinct randomized addition order parsimony starting trees (e.g., RAxML [Stamatakis 2014] or IQ-Tree (Nguyen et al., 2015)). While these search strategies are still highly likely to be stuck in local optima when

they converge, those local optima will have higher ML scores than the local optima the above simple algorithm will become stuck in.

As IQ-Tree mainly relies on NNI moves and is thus more prone to getting stuck in local optima, apart from using distinct parsimony starting trees, it deploys an additional technique. Once it is stuck in a so-called NNI optimum, it applies a couple of completely random NNI moves (without taking the likelihood score of these moves into account). This perturbs the tree sufficiently to move out of the local optimum. After this perturbation, a new round of NNI moves for improving the ML score is applied which is highly likely to end up in a distinct, potentially better, local optimum.

An alternative approach for navigating out of local optima by means of random perturbations is the ratchet method that has been applied to parsimony (Nixon, 1999) and ML searches (Vos, 2003). Here, the idea is to first randomly change the weights of the MSA sites, then apply a couple of NNI or SPR moves to this perturbed MSA, and subsequently restore the original MSA. Thereafter, one re-applies the search strategy to the new tree generated by the perturbed MSA.

GARLI (Zwickl, 2006) deploys a rather different approach to escaping local optima. It uses a so-called genetic algorithm. Instead of working on a single tree, GARLI conducts searches on a set (population) of trees. Periodically, information (e.g., subtrees) is exchanged among the trees in the population to improve the quality (ML scores) of the overall tree population.

Yet another option for escaping local optima is to deploy the simulated annealing search technique. Simulated annealing allows for carrying out backward steps, that is, it will occasionally conduct tree moves to trees with lower ML scores. This might allow to navigate out of local maxima in parsimony (Barker, 2004) and ML-based (Stamatakis, 2005) tree searches.

It is worth noting that simulated annealing is very similar to Bayesian MCMC sampling, as MCMC chains also occasionally accept so-called downhill steps to sample trees with a lower posterior probability.

Finally, there also exists the issue of what we have termed "rough likelihood surface". Such a rough likelihood surface typically emerges when analyzing datasets with comparatively few sites and a large number of taxa (e.g., a single-gene 16S RNA alignment comprising thousand taxa, or more). Typically, the search space will exhibit a large number of local optima that (i) can not be distinguished from each other using the standard likelihood-based significance tests as implemented, for instance, in CONSEL (Shimodaira and Hasegawa, 2001) and (ii) that exhibit large pairwise topological distances exceeding 20% or even 30% on average. Stamatakis (2011) gave an example of such a rough likelihood surface using an empirical single-gene dataset. In general, the key challenge with such datasets is that 100 distinct ML searches are likely to yield 100 topologically substantially different, but statistically indistinguishable trees.

3.1 Divide-and-Conquer Approaches

Thus far, we have briefly discussed how the most popular ML tree search strategies work. An alternative approach to designing phylogenetic search strategies is to deploy a divide-and-conquer strategy. Here, the idea is to initially divide the input sequences into disjoint or partially overlapping sets of closely related sequences (e.g., simply by clustering sequences or using a parsimony tree), then infer individual trees on those subsets, and finally merge these subtrees into one large comprehensive tree. The merging step can also be interpreted as a supertree reconstruction problem, provided that the sequence subsets overlap.

In general, all attempts to devise efficient divide-and-conquer tree search algorithms for ML have not been particularly successful with respect to speed and/or accuracy improvements over the aforementioned search methods that operate on a comprehensive tree. It turns out that all divide-and-conquer methods that have been devised so far, require an additional global optimization of the tree topology, the branch lengths, and the model parameters on the comprehensive tree to compete (with respect to accuracy) with the standard approaches described in Section 3. This requirement for additional global optimizations on a comprehensive tree also has a negative impact on the potential speed savings a divide-and-conquer approach could have.

While this is not properly understood yet, we suspect that global optimization on the comprehensive tree *is* required, because the information on branch lengths and model parameters that is propagated from an individual subtree has an important influence on the tree topology and branch lengths in the remaining subtrees. In other words, the subproblems (subtrees) we are attempting to solve do not appear to be sufficiently independent from each other to allow for applying a divide-and conquer approach. Some examples for divide-and-conquer or similar approaches with rather disappointing results have been shown by Roshan et al. (2004), Le Vinh et al. (2005) and Izquierdo-Carrasco et al. (2011).

A somewhat related approach is the quartet puzzling idea as implemented in Tree-Puzzle (Schmidt et al., 2002). It initially builds quartet trees (trees with 4 taxa) from the MSA and then puzzles them together into a comprehensive tree.

3.2 Terraces in Tree Space

A recently discovered phenomenon affecting likelihood-based phylogenetic inference (ML and BI) are terraces in tree space (Sanderson et al., 2011). A terrace is a, potentially large set, of topologically distinct tree topologies with *exactly* the same analytical likelihood score. Note that, numerical likelihood values might differ slightly because of roundoff error propagation.

Under likelihood, terraces may emerge for partitioned phylogenomic alignments when using unlinked branch length estimates. Branches are said to be unlinked, when we estimate a completely independent set of branch lengths for each partition (e.g., each gene) of the phylogenomic MSA (also known as supermatrix in this context). Terraces only emerge when branch lengths are unlinked, because the overall likelihood score of the entire concatenated MSA is the sum of the independently *optimized* per-partition likelihoods. In contrast to this, when branch lengths are not unlinked (i.e., scaled or joint branch length estimates), they can not be optimized independently for each partition separately. In other words, the partitions are somehow connected to each other via the shared (potentially scaled) branch length values which prevents the emergence of terraces.

If an a MSA (under an unlinked branch model) and respective partitioning scheme contains one or several terraces depends on the missing data pattern (Sanderson et al., 2011). Therefore, in the context of designing tree searches, one should avoid conducting tree moves that will just take the search to another tree located on the same terrace, or at least omit such redundant computations. Tree moves omitting unnecessary likelihood computations in a prototype RAxML implementation were pioneered by Stamatakis and Alachiotis (2010), essentially through implicit recognition of terraces, even before they were mathematically characterized in 2011. Later on, Chernomor presented work on omitting redundant computations for standard tree move operations using more elegant data structures than we did as well as a production-level implementation in IQ-Tree (Chernomor et al., 2015, 2016).

While the techniques implemented in RAxML and IQ-Tree allow for avoiding redundant

likelihood calculations on a terrace, they do not provide explicit mechanisms to move away from the terrace in tree space.

Terraces are not only relevant with respect to reducing the computational cost of ML searches. Their presence can also mislead downstream analyses.

For instance, the presence of terraces can severely bias bootstrap and Bayesian support values (Sanderson et al., 2014). Moreover, it is currently unknown how many published phylogenetic trees actually do reside on a terrace. The first study devoted to this topic (Dobrin et al., 2018) analyzed a collection of 26 large empirical phylogenomic datasets and showed that terraces are present in "nearly all datasets" and that "terraces found during bootstrap resampling reduced overall support".

One reason for this is that standard phylogenetic inference tools do not routinely assess if the final tree they generated resided on a terrace or not. However, there now exists a highly efficient C++ library (Biczok et al., 2018) for this purpose, that has already been integrated into RAxML-NG.

4 Computing Support Values

An important aspect when conducting empirical phylogenetic studies is the inference of support values on trees. The standard method is the non-parametric bootstrap (BS) as proposed by Felsenstein (1985) (but also see an interesting very recent modification of the phylogenetic bootstrap by Lemoine et al., 2018) The idea is to re-sample sites from the original MSA at random with replacement to assemble a set of 100 or more slightly perturbed BS replicate MSAs. One then applies the same algorithm as used for inferring the best-known ML tree on the original —unperturbed— MSA to each of those BS replicate MSAs. This yields a set of 100 (or more) BS trees that can subsequently be used to, either build a consensus tree, or map branch support values to the best-known ML tree inferred on the original MSA. Finally, a more elaborate approach to mapping BS support values onto a given tree, that also takes the size of the respective tree space into account, has recently been presented (Lemoine et al., 2018).

How many BS replicates are required is hard to determine, but it seems to heavily depend on the data at hand. For a criterion to determine the number of BS replicates, see Pattengale et al. (2010).

Evidently, the bootstrap procedure is computationally extremely expensive. Thus, there have been several attempts to devise faster and hence more approximate methods for inferring support values on phylogenies. It is important to note that, they represent "approximations of an approximation" as finding the optimal ML tree is NP-hard and the strategies presented in Section 3 already represent *ad hoc* heuristics. As such, using these fast methods for inferring support values can always, and perhaps rightfully so, be criticized by reviewers as being too approximate. From our point of view, despite having designed an approximate method ourselves, they do not capture that well, the search complexity of the problem that is associated to the vast tree search space, and the mere fact that we are already using heuristics. Thus, from our personal point of view, it is always best to conduct BS replicate searches using the standard search strategies with the standard bootstrap procedure, if the dataset size and the available computational resources allow for this.

In RAxML we introduced the so-called rapid bootstrap (Stamatakis et al., 2008) that essentially relies on a more approximate, less thorough version of the standard RAxML search algorithm, that can more easily be trapped in local maxima.

The so-called ultrafast bootstrap (Minh et al., 2013) implemented in IQ-Tree relies on an

approximate sampling of emulated BS replicates during the search on the original MSA. As such, it depends heavily on whether the regions of tree space that are explored by the search on the original MSA also form part of the BS replicate tree search space. In other words, one needs those two regions (original MSA and BS replicate MSA region) of the tree search space to overlap sufficiently in order for the approximation to be accurate. This is, however, not guaranteed a priori.

Finally, the approximate Likelihood Ratio Test (aLRT, Anisimova and Gascuel, 2006) takes a given best-known ML tree and conducts a statistical test on each inner branch of the tree by computing likelihood scores for the three possible NNI configurations (see Figure 4) around this branch and subsequently using them for the test statistic. The aLRT can be criticized for exclusively relying on the very local NNI-based likelihoods for computing the test statistics.

We believe that all of the above approximations for obtaining support values are useful. We nonetheless wish to emphasize that they remain approximations of an approximation and are thus prone to criticism.

5 Conclusion

In this book chapter we have presented the basic components, tree alteration operations, and flavors of commonly used tree search strategies under ML. Moreover, we discussed the standard phylogenetic BS procedure for inferring support values on trees as well as some faster and more approximate methods for this task. We also briefly reviewed the time complexity for finding *the* ML tree and explained the intuition behind the concept of NP-hardness. Finally, we also briefly outlined the additional problems that arise when the dataset to be analyzed contains so-called terraces in its tree search space.

Practitioners should keep in mind that all tree search tools implement *ad hoc* heuristic search strategies that have been developed and tested using some simulated and some empirical benchmark test datasets. It is thus likely that they will fail, that is, perform sub-optimally on other datasets under distinct or difficult settings. Also, the search strategies presented here do not have any performance guarantees, that is, how far away from the global maximum the trees they infer are.

The best approach is to conduct searches using several ML tree inference tools (e.g. Chapter 1.3 [Kozlov and Stamatakis 2020]) as well as tools for Bayesian phylogenetic inference (e.g. Chapter 1.5 [Lartillot 2020b]) and subsequently compare the results. This also helps to minimize the impact of potential programming errors (Darriba et al., 2018) as tree inference software has become more complex and supports substantially more, as well as more complex models than 10 - 15 years ago.

Acknowledgements

The authors wish to thank the following former students of our 2018 summer school on computational molecular evolution for useful comments on the initial draft of this book chapter: Paschalis Natsidis and Alexandros Vasilikopoulos.

References

Aberer, A. J., Kobert, K., and Stamatakis, A. (2014). Exabayes: Massively parallel bayesian tree inference for the whole-genome era. *Molecular biology and evolution*, 31(10):2553–2556.

Anisimova, M. and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic biology*, 55(4):539–552.

Barker, D. (2004). Lvb: parsimony and simulated annealing in the search for phylogenetic trees. *Bioinformatics (Oxford, England)*, 20(2):274–275.

Biczok, R., Bozsoky, P., Eisenmann, P., Ernst, J., Ribizel, T., Scholz, F., Trefzer, A., Weber, F., Hamann, M., and Stamatakis, A. (2018). Two c++ libraries for counting trees on a phylogenetic terrace. *Bioinformatics*.

Chernomor, O., Minh, B. Q., and von Haeseler, A. (2015). Consequences of common topological rearrangements for partition trees in phylogenomic inference. *Journal of Computational Biology*, 22(12):1129–1142.

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic biology*, 65(6):997–1008.

Darriba, D., Flouri, T., and Stamatakis, A. (2018). The state of software for evolutionary biology. *Molecular biology and evolution*, 35(5):1037–1046.

Day, W. H. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49(4):461–467.

Dobrin, B. H., Zwickl, D. J., and Sanderson, M. J. (2018). The prevalence of terraced treescapes in analyses of phylogenetic data sets. *BMC evolutionary biology*, 18(1):46.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791.

Gascuel, O. (1997). Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Molecular biology and evolution*, 14(7):685–695.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Systematic biology*, 59(3):307–321.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5):696–704.

Izquierdo-Carrasco, F., Smith, S. A., and Stamatakis, A. (2011). Algorithms, data structures, and numerics for likelihood-based phylogenetic inference of huge trees. *BMC bioinformatics*, 12(1):470.

Kozlov, A. M. and Stamatakis, A. (2020). Using raxml-ng in practice. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.3, pages 1.3:1–1.3:25. No commercial publisher | Authors open access book.

Lartillot, N. (2020a). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. (2020b). Phylobayes: Bayesian phylogenetics using site-heterogeneous models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.5, pages 1.5:1–1.5:16. No commercial publisher | Authors open access book.

Le Vinh, S., Schmidt, H. A., and von Haeseler, A. (2005). Phynav: A novel approach to reconstruct large phylogenies. In *Classification—the Ubiquitous Challenge*, pages 386–393. Springer.

Lemoine, F., Entfellner, J.-B. D., Wilkinson, E., Correia, D., Felipe, M. D., Oliveira, T., and Gascuel, O. (2018). Renewing felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, 556(7702):452.

Minh, B. Q., Nguyen, M. A. T., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Molecular biology and evolution*, 30(5):1188–1195.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763–767.

Morrison, D. A. (2007). Increasing the Efficiency of Searches for the Maximum Likelihood Tree in a Phylogenetic Analysis of up to 150 Nucleotide Sequences. *Systematic Biology*, 56(6):988–1010.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274.

Nixon, K. C. (1999). The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15(4):407–414.

Nylander, J. A., Wilgenbusch, J. C., Warren, D. L., and Swofford, D. L. (2008). Awty (are we there yet?): a system for graphical exploration of mcmc convergence in bayesian phylogenetics. *Bioinformatics*, 24(4):581–583.

Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. (1994). fastdnaml: a tool for construction of phylogenetic trees of dna sequences using maximum likelihood. *Bioinformatics*, 10(1):41–48.

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M., and Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of computational biology*, 17(3):337–354.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147.

Roch, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 3(1):92.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542.

Roshan, U., Moret, B. M., Williams, T. L., and Warnow, T. (2004). Rec-i-dcm3: A fast algorithmic technique for reconstructing large phylogenetic trees. In *Proc. 3rd IEEE Computational Systems Bioinformatics Conf. CSB" 04*, LCBB-CONF-2004-002, pages 98–109. IEEE Press.

Sanderson, M. J., McMahon, M. M., Stamatakis, A., Zwickl, D. J., and Steel, M. (2014). Impacts of terraces on phylogenetic inference. *arXiv preprint arXiv:1410.8071*.

Sanderson, M. J., McMahon, M. M., and Steel, M. (2011). Terraces in phylogenetic tree space. *Science*, 333(6041):448–450.

Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3):502–504.

Shimodaira, H. and Hasegawa, M. (2001). Consel: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17(12):1246–1247.

Stamatakis, A. (2005). An efficient program for phylogenetic inference using simulated annealing. In *Parallel and Distributed Processing Symposium, 2005. Proceedings. 19th IEEE International*, pages 8–pp. IEEE.

Stamatakis, A. (2011). Phylogenetic search algorithms for maximum likelihood. *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*, pages 547–577.

1.2:18 REFERENCES

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stamatakis, A. and Alachiotis, N. (2010). Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics*, 26(12):i132–i139.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the raxml web servers. *Systematic biology*, 57(5):758–771.

Swofford, D. L. (2001). Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0. b5.

Vos, R. (2003). Accelerated likelihood surface exploration: the likelihood ratchet. *Systematic Biology*, 52(3):368–373.

Zwickl, D. (2006). *GARLI, vers. 0.951. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.* PhD thesis, Ph. D. dissertation, University of Texas, Austin, Texas, USA.

Chapter 1.3 Using RAxML-NG in Practice

Alexey M. Kozlov
Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
Alexey.Kozlov@h-its.org
http://orcid.org/0000-0001-7394-2718

Alexandros Stamatakis
Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies,
Karlsruhe Institute of Technology, Institute for Theoretical Informatics
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
Alexandros.Stamatakis@h-its.org
https://orcid.org/0000-0003-0353-0691

Abstract

RAxML-NG is a new phylogenetic inference tool that replaces the widely-used RAxML and ExaML tree inference codes. Compared to its predecessors, RAxML-NG offers improvements in accuracy, flexibility, speed, scalability, and user-friendliness. In this chapter, we provide practical recommendations for the most common use cases of RAxML-NG: tree inference, branch support estimation via non-parametric bootstrapping, and parameter optimization on a fixed tree topology. We also describe best practices for achieving optimal performance with RAxML-NG, in particular, with respect to parallel tree inferences on computer clusters and supercomputers. As RAxML-NG is continuously updated, the most up-to-date version of the tutorial described in this chapter is available online at: `https://cme.h-its.org/exelixis/raxml-ng/tutorial`

How to cite: Alexey M. Kozlov and Alexandros Stamatakis (2020). Using RAxML-NG in Practice. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 1.3, pp. 1.3:1–1.3:25. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Supplement Material https://github.com/amkozlov/ng-tutorial

Funding This work was financially supported by the Klaus Tschira Foundation.

1 Introduction

RAxML (Stamatakis, 2006b, 2014) is a widely-used tool for maximum likelihood (ML) based phylogenetic inference (see Chapter 1.2 [Stamatakis and Kozlov 2020]). It has been cited by more than 25,000 publications over the last 15 years. More recently, we introduced ExaML (Stamatakis and Aberer, 2013; Kozlov et al., 2015), a variant of RAxML with several novel features including checkpointing, improved load balancing, and an efficient fine-grained MPI parallelization. These improvements were particularly important for being able to analyze large-scale phylogenomic datasets on compute clusters and supercomputer systems (Misof et al., 2014; Jarvis et al., 2014). However, ExaML only offered a core subset of RAxML functionalities. It lacks several important functions such as bootstrapping and comprehensive starting tree generation. These limitations, and its dependency on MPI, made ExaML more difficult to install and use, and therefore presumably limited its adoption.

1.3:2 RAxML-NG

With RAxML-NG, we introduced one single code that scales from the laptop to the supercomputer. It combines the parallel efficiency of ExaML with the functional completeness of RAxML. Furthermore, RAxML-NG is more user-friendly than RAxML/ExaML because of a simplified installation process, the re-engineered command line interface, and new default settings that cover the most common usage scenarios (see Appendix A).

RAxML-NG can be downloaded at `https://github.com/amkozlov/raxml-ng`. The corresponding documentation is available via a `GitHub` wiki at `https://github.com/amkozlov/raxml-ng/wiki`. Technical implementation details and benchmarking results are provided by Kozlov et al. (2019) and in Chapter 4 in Kozlov (2018). We also offer extensive user support via the RAxML google group: `https://groups.google.com/forum/#!forum/raxml`.

All datasets used in this chapter can be downloaded from `https://github.com/amkozlov/ng-tutorial`.

IMPORTANT NOTE: You will need RAxML-NG 0.8.0b or later for this tutorial, so please make sure you have the right version:

```
$ raxml-ng -v

RAxML-NG v. 0.8.0 BETA released on 11.01.2019 by The Exelixis Lab.
```

2 Pre-processing the alignment

2.1 Sanity check

Before starting the actual analysis, it is strongly recommended to perform a multiple sequence alignment (MSA) sanity check by calling RAxML-NG with the `--check` option.

```
$ raxml-ng --check --msa bad.fa --model GTR+G
```

This command will check the MSA for several common format issues as well as data inconsistencies including:

- duplicate taxon names
- invalid characters in taxon names
- duplicate sequences
- fully undetermined ("gap-only") sequences and columns
- incorrect or incompatible evolutionary models, partitioning scheme and starting trees (if provided)

Performing this check before starting the analysis is very important, since based on our experience, a large proportion of failed RAxML runs are due to tree or MSA format errors!

Let us take a closer look at the output of our sanity check invocation:

```
WARNING: Fully undetermined columns found: 2

WARNING: Fully undetermined sequences found: 2

WARNING: Sequences t3 1200bp and t8 are exactly identical!
WARNING: Duplicate sequences found: 1

ERROR: Following taxon name contains invalid characters: t9'
ERROR: Following taxon name contains invalid characters: t6)
```

```
ERROR: Following taxon name contains invalid characters: t3 1200bp

ERROR: Alignment check failed (see details above)!
```

It seems that this MSA file has almost every conceivable problem. RAxML-NG will attempt to automatically fix the most common issues, for instance, by removing fully undetermined columns and sequences, replacing invalid characters in taxon names etc. To achieve this, it will write an analogously updated/fixed MSA file back to disk:

```
NOTE: Reduced alignment (with duplicates and gap-only sites/taxa removed)
NOTE: was saved to: /home/alexey/ng-tutorial/bad.fa.raxml.reduced.phy
```

Let us now repeat the sanity check with the fixed file:

```
$ raxml-ng --check --msa bad.fa.raxml.reduced.phy --model GTR+G

[..]
Alignment can be successfully read by RAxML-NG.
```

2.2 Compression and conversion to binary format

For large alignments, we recommend using the `--parse` command after, or, instead of `--check`:

```
$ raxml-ng --parse --msa prim.phy --model GTR+G
```

In addition to the MSA sanity check, this command will compress alignment patterns and store the MSA in a binary format (RAxML Binary Alignment, RBA):

```
NOTE: Binary MSA file created: prim.phy.raxml.rba
```

In the process of pattern compression, RAxML-NG identifies identical MSA sites and converts them into a single site ('pattern') with a weight corresponding to their number of occurrences. Since this compression step can potentially require quite some time for broad supermatrix MSAs, directly loading a RBA file is (substantially) faster compared to parsing and loading a plain FASTA or PHYLIP file. This parsing speed is important for large-scale parallel tree inferences with say, 500 cores or more, as virtually no time is lost at the beginning for parsing the file and the cores can almost immediately start with the likelihood calculations (see Kozlov et al., 2015, Supplement Section 3 for more details).

In addition, `--parse` will estimate the memory requirements and optimal number of CPUs/threads for the particular MSA:

```
* Estimated memory requirements                : 2 MB
* Recommended number of threads / MPI processes: 2
```

Even though these estimates are approximate, they provide a "good" starting point for experimentation (see Section 7 for details) to determine the optimal number of cores that will yield maximum parallel efficiency.

3 Inferring ML trees

Let us now infer a tree under the `GTR+GAMMA` (general time reversible model of nucleotide substitution with a Γ model of rate heterogeneity) model with default parameters. We

will use 2 threads as suggested above, and provide a fixed random number seed to ensure reproducibility. By using a fixed random number seed RAxML-NG will always produce the same sequence of random numbers and therefore a failed run can be easily reproduced for debugging. Note that, we will also always use a new name via the `--prefix` output file name option for each RAxML-NG example run to avoid overwriting preceding output files.

```
$ raxml-ng --msa prim.phy --model GTR+G --prefix T3 --threads 2 --seed 2
```

The above command will perform 20 tree searches using 10 random and 10 parsimony-based starting trees. In the end it will pick the best-scoring topology:

```
Analysis options:
  run mode: ML tree search
  start tree(s): random (10) + parsimony (10)
```

This default setting represents a reasonable choice for most practical cases. However, computational resources permitting, we might want to increase the number of starting trees to explore the tree space more thoroughly:

```
$ raxml-ng --msa prim.phy --model GTR+G --prefix T4 --threads 2 --seed 2
    --tree pars{25},rand{25}
```

Conversely, we can also just perform a quick-and-dirty search from a single random starting tree using the `--search1` command:

```
$ raxml-ng --search1 --msa prim.phy --model GTR+G --prefix T5 --threads 2
    --seed 2
```

Let us now compare the results of all three alternative tree inference runs:

```
$ grep "Final LogLikelihood:" T{3,4,5}.raxml.log

T3.raxml.log:Final LogLikelihood: -5708.923977
T4.raxml.log:Final LogLikelihood: -5708.923977
T5.raxml.log:Final LogLikelihood: -5708.979717
```

This looks quite good: the likelihood surface appears to have a clear peak, which RAxML-NG finds regardless of the search parameters (up to a small deviation in "T5", which we will explain below).

We use the term *likelihood surface* in a colloquial/subjective way to describe the space of all possible tree topologies and their respective likelihood scores. If the likelihood surface is smooth there seems to be one clear peak that is identified by several independent searches. If the surface is rough, we typically observe a plethora of substantially different tree topologies but with statistically indistinguishable likelihood scores. Rough likelihood surfaces are frequently observed for large single gene MSAs with $1,000$ or more sequences (see Section 3.2 in Chapter 1.2 [Stamatakis and Kozlov 2020]).

Let us get back to our example. We observe that the tree "T5" has a slightly worse likelihood. The question arises if it also has a distinct topology. We can check this by using the `--rfdist` command to compute the topological Robinson-Foulds (RF) distance (Robinson and Foulds, 1981) between all trees we have inferred:

```
$ cat T{3,4}.raxml.mlTrees T5.raxml.bestTree > mltrees
$ raxml-ng --rfdist --tree mltrees --prefix RF
```

```
[...]
Loaded 71 trees with 12 taxa.

Average absolute RF distance in this tree set: 0.000000
Average relative RF distance in this tree set: 0.000000
Number of unique topologies in this tree set: 1
```

This tells us that, in fact, all 71 resulting topologies (one per starting tree) are identical, so we can be optimistic that we found *the* globally optimal ML tree. The slight numerical deviations we observe for the likelihood scores are due to numerical round-off error propagation. To conduct calculations computers rely on so-called floating-point numbers that are just an imperfect representation of the real numbers on the machine.

Unfortunately, not all datasets are as well-behaved as our initial test dataset:

```
$ raxml-ng --msa fusob.phy --model GTR+G --prefix T6 --seed 2 --threads 2
$ grep "ML tree search #" T6.raxml.log

[00:00:03] ML tree search #1, logLikelihood: -9974.668088
[00:00:07] ML tree search #2, logLikelihood: -9974.666644
[00:00:11] ML tree search #3, logLikelihood: -9974.669417
[00:00:15] ML tree search #4, logLikelihood: -9974.664855
[00:00:19] ML tree search #5, logLikelihood: -9974.663779
[00:00:22] ML tree search #6, logLikelihood: -9974.666906
[00:00:26] ML tree search #7, logLikelihood: -9974.668155
[00:00:30] ML tree search #8, logLikelihood: -9974.664340
[00:00:33] ML tree search #9, logLikelihood: -9974.666937
[00:00:37] ML tree search #10, logLikelihood: -9974.666388
[00:00:40] ML tree search #11, logLikelihood: -9980.601114
[00:00:43] ML tree search #12, logLikelihood: -9974.675123
[00:00:46] ML tree search #13, logLikelihood: -9980.602470
[00:00:49] ML tree search #14, logLikelihood: -9974.671637
[00:00:52] ML tree search #15, logLikelihood: -9980.602668
[00:00:54] ML tree search #16, logLikelihood: -9980.601182
[00:00:57] ML tree search #17, logLikelihood: -9974.672801
[00:01:00] ML tree search #18, logLikelihood: -9974.668668
[00:01:03] ML tree search #19, logLikelihood: -9974.669997
[00:01:06] ML tree search #20, logLikelihood: -9980.607281
```

This example illustrates why it is so important to use multiple starting trees: we can see that some searches ended up in a local optimum with a substantially lower likelihood (-9980.607281 vs. -9974.669997). Once again, let us check if the resulting trees differ topologically:

```
$ raxml-ng --rfdist --tree T6.raxml.mlTrees --prefix RF6
[...]
Loaded 20 trees with 38 taxa.

Average absolute RF distance in this tree set: 3.157895
Average relative RF distance in this tree set: 0.045113
```

```
Number of unique topologies in this tree set: 2

Pairwise RF distances saved to: <...>/RF6.raxml.rfDistances
```

So we have 2 distinct topologies in our set of 20 inferred trees, which correspond to two distinct likelihood values we observed in the tree search output. Let us look at the individual pairwise RF distances which are printed to the `RF6.raxml.rfDistances` file:

```
$ cat RF6.raxml.rfDistances

0       1       0       0.000000
0       2       0       0.000000
0       3       0       0.000000
0       4       0       0.000000
0       5       0       0.000000
0       6       0       0.000000
0       7       0       0.000000
0       8       0       0.000000
0       9       0       0.000000
0       10      8       0.114286
0       11      0       0.000000
0       12      8       0.114286
0       13      0       0.000000
0       14      8       0.114286
0       15      8       0.114286
0       16      0       0.000000
0       17      0       0.000000
0       18      0       0.000000
0       19      8       0.114286
[...]
```

Here, 1st and 2nd column contain tree indices in the NEWICK file, 3rd column shows the *absolute* RF distance between those two trees, and 4th column shows the *relative* or *normalized* RF distance ranging from 0 to 1. Relative RF distance is calculated by dividing the absolute value (column 3) by the maximum possible RF distance for this tree pair.

As we can see, all 10 searches from the random starting trees (trees 0 to 9) found the best-scoring topology (RF=0, logL=−9974), whereas 5 out of 10 searches from a parsimony starting tree converged to a local optimum (RF = 8, logL = −9980). Ideally, one should also check whether the likelihood difference between both topologies is statistically significant. This could be done, for example, by CONSEL (Shimodaira and Hasegawa, 2001), which implements a large number of statistical significance tests for comparing tree topologies.

4 Bootstrapping and branch support

NOTE: As of v.0.8.0*b*, RAxML-NG only supports the *standard* bootstrap algorithm (corresponding to the `-b` option in standard RAxML). It is subsantially slower than *rapid* bootstrapping implemented in standard RAxML (`-x` or `-f a` options), but returns more accurate support values.

4.1 Inferring bootstrap trees

RAxML-NG can perform the standard non-parametric bootstrap by re-sampling alignment columns and re-inferring a tree for each bootstrap (BS) replicate MSA:

```
raxml-ng --bootstrap --msa prim.phy --model GTR+G --prefix T7 --seed 2 --
    threads 2
```

By default, RAxML-NG employs the so-called MRE-based bootstopping test (Pattengale et al., 2010) to automatically determine the sufficient number of BS replicates. The diagnostic statistics is evaluated after every 50 BS tree inferences, and once its value drops below the cutoff, the analysis stops. The key motivation for the bootstopping criterion is to ensure that neither too few (unstable/inaccurate support values) nor too many (waste of CPU time) replicates are computed. To assess stability of support values, the bootstopping criterion repeatedly splits the current set of BS replicate trees at random into two tree sets of equal size and subsequently compares the support values induced by these sets. If the induced support values are not substantially different it suggests that bootstrapping should stop.

Let us now infer some BS replicates:

```
  bootstrap replicates: max: 1000 + bootstopping (autoMRE, cutoff:
    0.030000)
[...]
[00:00:15] Bootstrap tree #50, logLikelihood: -5762.777409
[00:00:15] Bootstrapping converged after 50 replicates.
```

This converged quickly! Let us now manually increase the number of BS replicates to be on the safe side:

```
raxml-ng --bootstrap --msa prim.phy --model GTR+G --prefix T8 --seed 2 --
    threads 2 --bs-trees 200
```

Bootstrap convergence can also be assessed *after* the BS inference via the `--bsconverge` command. Note that, we can also change the bootstopping cutoff value to make the test more or less stringent:

```
$ raxml-ng --bsconverge --bs-trees T7.raxml.bootstraps --prefix T9 --seed
    2 --threads 2 --bs-cutoff 0.01

# trees   avg WRF  avg WRF in %   # perms: wrf <= 1.00 %  converged?
     50    7.400          1.644                        0     NO
Bootstopping test did not converge after 50 trees
```

The cutoff here represents the bipartition-frequency-weighted RF distance (WRF, see Pattengale et al., 2010, for details) between extended majority rule consensus trees calculated on the respective randomly split BS tree set. By default we calculate 1000 such random splits of the tree set and average the WRF distances over them.

As we can see, with a 1% WRF cutoff 50 replicates are not enough. What about 200?

```
$ raxml-ng --bsconverge --bs-trees T8.raxml.bootstraps --prefix T10 --seed
     2 --threads 2  --bs-cutoff 0.01

# trees   avg WRF  avg WRF in %   # perms: wrf <= 1.00 %  converged?
     50    7.400          1.644                        0     NO
```

```
    100   11.702          1.300                       245     NO
    150   13.960          1.034                       457     NO
    200   16.484          0.916                       648     NO
Bootstopping test did not converge after 200 trees
```

Still no convergence, but the WRF distance (`avg WRF in %`) is steadily decreasing as we add more replicates, and now lies below the 1% cutoff for 648 out of the 1000 random splits of the BS tree set (convergence requirement: > 990). This looks promising, and we can expect convergence after few hundred replicates. Luckily, bootstraps are independent, and we can thus reuse the 200 BS trees we have already inferred. So let's add 400 additional BS replicate trees.

IMPORTANT NOTE: It is extremely important to specify a distinct random seed for the second run, otherwise first 200 trees of the second run will be identical to the first run!

```
raxml-ng --bootstrap --msa prim.phy --model GTR+G --prefix T11 --seed 333
    --threads 2 --bs-trees 400
```

Now, we can simply concatenate the BS replicate trees from both runs, and re-assess the convergence:

```
$ cat T8.raxml.bootstraps T11.raxml.bootstraps > allbootstraps
$ raxml-ng --bsconverge --bs-trees allbootstraps --prefix T12 --seed 2 --
    threads 1 --bs-cutoff 0.01

# trees   avg WRF  avg WRF in %   # perms: wrf <= 1.00 %  converged?
     50    7.400          1.644                        0     NO
     100  11.702          1.300                      245     NO
     150  13.960          1.034                      457     NO
     200  16.484          0.916                      648     NO
     250  17.410          0.774                      841     NO
     300  18.900          0.700                      927     NO
     350  20.060          0.637                      942     NO
     400  22.076          0.613                      969     NO
     450  23.856          0.589                      973     NO
     500  26.164          0.581                      985     NO
     550  27.844          0.563                      985     NO
     600  28.462          0.527                      991     YES
Bootstopping test converged after 600 trees
```

Now we have convergence, even with a more stringent bootstopping cutoff. However, we had to conduct 600 BS replicate searches instead of just 50. On large datasets, this quickly becomes computationally expensive. Hence in practice, the default bootstopping cutoff value of an average WRF of 3% should be sufficient in most cases (Pattengale et al., 2010).

4.2 Computing branch support

Now, what can we do with the BS trees? We can either summarize them via some sort of consensus tree (strict, majority, majority rule extended, e.g., using the `--consense` command of RAxML-NG) or we can map them onto the best-scoring ML tree that we inferred on the original MSA. It is debatable what the best way of summarizing BS trees might be, but there seems to be a trend toward mapping the BS support values onto the best-scoring/best-known

ML tree (remember: finding *the* globally optimal ML tree is computationally hard), so let us do that.

We will use the ML tree obtained in run `T3` (see Section 3):

```
raxml-ng --support --tree T3.raxml.bestTree --bs-trees allbootstraps --
    prefix T13 --threads 2
```

Now, we can actually look at this best-known ML tree including supports, contained in `T13.raxml.support`, using some tree viewer (e.g., Dendroscope [Huson and Scornavacca 2012] or FigTree [Rambaut 2012]). **Beware**: due to confusion between *node* and *branch* attributes in the NEWICK format, some viewers have or had issues concerning correct branch support visualization (Czech et al., 2017). If possible (e.g., in recent versions of Dendroscope), you should specify that support values must be interpreted as *edge* labels.

Alternatively, we can also compute the so-called *Transfer Bootstrap Expectation* (*TBE*) support metric recently suggested by Lemoine et al. (2018) as follows:

```
$ raxml-ng --support --tree T3.raxml.bestTree --bs-trees allbootstraps --
    prefix T14 --threads 2 --bs-metric tbe
```

While the standard bootstrap support metric (*Felsenstein's bootstrap*, *FBP*) relies on binary presence/absence of bipartitions from replicate trees in the best-known ML tree, TBE is based on a gradual 'transfer' distance. Transfer distance between two branches equals to the minimum number of taxa that have to be transfered (or removed) to make those branches identical (that is, both branches split the set of taxa in identical subsets). TBE support for a branch in the ML tree is computed based on the *minimum* transfer distance between this branch and *any* branch in the BS replicate tree; in other words, we compare each ML tree branch to its respective closest branch in the BS replicate tree (please see Lemoine et al., 2018, for details). For this reason, TBE can better recover support in very large trees with thousands of taxa. This is because, bipartitions that exactly match those in the best-known ML tree are rarely present in replicates, and thus FBP usually yields low support, especially for deep branches.

As shown above, TBE can be computed from the same set of bootstrap replicate trees, so there in no need to repeat the compute-intensive tree inference step. However, the TBE computation itself is more expensive than FBP. This can be noted when computing the TBE on large trees: e.g., on a laptop, RAxML-NG v0.8.0 needs ≈ 20 seconds per BS replicate tree on the $9,000$ taxon dataset by Lemoine et al. (2018). However, this time is still negligible compared to the time required for BS replicate tree inference.

Finally, RAxML-NG offers a convenient "all-in-one" analysis mode for really lazy users (analogous to `-f a` in standard RAxML):

```
$ raxml-ng --all --msa prim.phy --model GTR+G --prefix T15 --seed 2 --
    threads 2 --bs-metric fbp,tbe
```

This will do all of the above steps (20 ML inferences on the original MSA, inferring bootstrap replicate trees, applying MRE-based bootstopping test, and drawing support values using both FBP and TBE on the best-scoring tree) with just a single command:

```
$ ls T15.*
T15.raxml.bestModel  T15.raxml.bestTree
T15.raxml.bootstraps T15.raxml.log
T15.raxml.mlTrees    T15.raxml.rba
T15.raxml.startTree  T15.raxml.supportFBP
```

```
T15.raxml.supportTBE
```

Please note, that for taxa-rich alignments running such a complete analysis with the `--all` command can take extremely long. It is therefore recommended to estimate the runtime required for a single tree search first, for instance, by using the `--search1` command. Based on the results, one might consider allocating more CPU cores and/or using the coarse-grained parallelization (see Section 7.7).

5 Tree likelihood evaluation

5.1 Basics

Another standard task is to evaluate trees, that is, to compute the likelihood of a given fixed tree topology by just optimizing model and/or branch length parameters on that fixed tree. This operation is frequently needed in model and hypothesis testing.

The basic option is `--evaluate`. It will re-optimize all branch lengths and all free model parameters. This default behavior can be altered with `--opt-branches on/off` and `--opt-model on/off`. There is also the `--loglh` command which is a short alias for

```
--evaluate --opt-branches off --opt-model off --nofiles
```

that is, it will compute and print the likelihood of the tree(s) without optimizing anything and without creating any output files. For instance, we can re-compute the likelihood of `T3` with *default* model parameters as follows:

```
$ raxml-ng --loglh --msa prim.phy --model GTR+G --tree T3.raxml.bestTree
    --threads 2

   Rate heterogeneity: GAMMA (4 cats, mean),  alpha: 1.000000 (ML),
    weights&rates: (0.250000,0.136954) (0.250000,0.476752)
    (0.250000,1.000000) (0.250000,2.386294)
   Base frequencies (ML): 0.250000 0.250000 0.250000 0.250000
   Substitution rates (ML): 1.000000 1.000000 1.000000 1.000000 1.000000
    1.000000

Final LogLikelihood: -6420.095053
```

In contrast, after re-optimizing all model parameters we obtain:

```
$ raxml-ng --evaluate --msa prim.phy --model GTR+G --tree T3.raxml.
    bestTree --threads 2 --nofiles

   Rate heterogeneity: GAMMA (4 cats, mean),  alpha: 0.377068 (ML),
    weights&rates: (0.250000,0.013550) (0.250000,0.164429)
    (0.250000,0.705224) (0.250000,3.116797)
   Base frequencies (ML): 0.354236 0.321458 0.080986 0.243320
   Substitution rates (ML): 3.989744 45.320369 3.326172 2.533579 36.939966
     1.000000

Final LogLikelihood: -5709.002997
```

Finally, we can fix some parameters to certain values and optimize others:

```
$ raxml-ng --evaluate --msa prim.phy --model GTR+G{2.0}+F{0.2/0.3/0.4/0.1}
     --tree T3.raxml.bestTree --threads 2 --nofiles

   Rate heterogeneity: GAMMA (4 cats, mean),  alpha: 2.000000 (user),
    weights&rates: (0.250000,0.293275) (0.250000,0.655014)
    (0.250000,1.069990) (0.250000,1.981722)
   Base frequencies (user): 0.200000 0.300000 0.400000 0.100000
   Substitution rates (ML): 44.454379 47.979464 65.161744 2.413970
    252.745302 1.000000

Final LogLikelihood: -6158.335994
```

Full list of supported evolutionary models and respective parameters can be found at https://github.com/amkozlov/raxml-ng/wiki/Input-data#single-model.

5.2 Comparing different models

Let us now conduct some small tests that show how the likelihood improves as we add more and more free parameters to our model. For this, we will use the best-scoring ML tree from Section 3 again.

Let us first evaluate the tree under the most simple model, Jukes-Cantor (`JC`):

```
$ raxml-ng --evaluate --msa prim.phy --threads 2 --model JC --tree T3.
    raxml.bestTree --prefix E1
```

Now, let us add the Γ model of rate heterogeneity:

```
$ raxml-ng --evaluate --msa prim.phy --threads 2 --model JC+G --tree T3.
    raxml.bestTree --prefix E2
```

Now let us use a simple `GTR` model (without rate heterogeneity):

```
$ raxml-ng --evaluate --msa prim.phy --threads 2 --model GTR --tree T3.
    raxml.bestTree --prefix E3
```

`GTR` with the `GAMMA` model of rate heterogeneity, but using empirical base frequencies:

```
$ raxml-ng --evaluate --msa prim.phy --threads 2 --model GTR+G+FC --tree
    T3.raxml.bestTree --prefix E4
```

And now also conducting a ML estimate of the base frequencies:

```
$ raxml-ng --evaluate --msa prim.phy --threads 2 --model GTR+G+FO --tree
    T3.raxml.bestTree --prefix E5
```

Finally, using 4 free rates (Yang, 1995) instead of `GAMMA`-distributed rates:

```
$ raxml-ng --evaluate --msa prim.phy --threads 2 --model GTR+R4+FO --tree
    T3.raxml.bestTree --prefix E6
```

Let us check the results:

```
$ grep logLikelihood E*.raxml.log

E1.raxml.log:[00:00:00] Tree #1, final logLikelihood: -6424.203056 <- JC
E2.raxml.log:[00:00:00] Tree #1, final logLikelihood: -6272.469063 <- JC+
    GAMMA
E3.raxml.log:[00:00:00] Tree #1, final logLikelihood: -5934.158958 <- GTR
E4.raxml.log:[00:00:00] Tree #1, final logLikelihood: -5719.408956 <- GTR
    + GAMMA + empirical base freqs
E5.raxml.log:[00:00:00] Tree #1, final logLikelihood: -5709.002997 <- GTR
    + GAMMA + estimated base freqs
E6.raxml.log:[00:00:01] Tree #1, final logLikelihood: -5706.008654 <- GTR
    + FreeRate + estimated base freqs
```

Unsurprisingly, models with more free parameters yield better likelihood scores. However, this does not mean that we should always use the most parameter-rich model. Instead, it is common to use information theoretical criteria such as `AIC` (Akaike Information Criterion), `AICc` (corrected AIC; AIC with a correction for small sample sizes) or `BIC` (Bayesian Information Criterion) to penalize parameter-rich models and thereby avoid overfitting the data. The three aforementioned criteria are implemented in RAxML-NG:

```
$ grep "AIC score" E*.raxml.log

E1.raxml.log:AIC score: 12890.406112 / AICc score: 12891.460907 / BIC
    score: 12991.209684  <- JC
E2.raxml.log:AIC score: 12588.938126 / AICc score: 12590.094698 / BIC
    score: 12694.541868  <- JC+G
E3.raxml.log:AIC score: 11926.317917 / AICc score: 11928.322525 / BIC
    score: 12065.522849  <- GTR
E4.raxml.log:AIC score: 11498.817912 / AICc score: 11500.963241 / BIC
    score: 11642.823014  <- GTR+G+FC
E5.raxml.log:AIC score: 11478.005995 / AICc score: 11480.151323 / BIC
    score: 11622.011097  <- GTR+G+FO
E6.raxml.log:AIC score: 11482.017308 / AICc score: 11484.940742 / BIC
    score: 11650.023260  <- GTR+R4+FO
```

For all criteria, model with the lowest score should be preferred. As we can see, the `GTR+G+FO` model scores best according to all three information theoretical criteria evaluated, even though it yields a lower likelihood than `GTR+R4+FO`. This example illustrates the importance of formal model selection. In practice, one should use specialized tools such as ModelTest-NG (Darriba et al., 2020), IQTree/ModelFinder (Kalyaanamoorthy et al., 2017), or PartitionFinder (Lanfear et al., 2016) for this task.

6 Partitioned analyses

6.1 Partitioned model definition

So far, we always used a single evolutionary model for all MSA sites. This is biologically rather unrealistic, since different genes and/or codon positions typically exhibit distinct substitution patterns. Therefore, it is common to divide MSA sites into subsets or *partitions*,

to which we can assign individual evolutionary models. In the most simple case, we can assign identical *models* to all partitions, but allow for independent *model parameter* estimates:

```
$ cat prim.part

GTR+G+FO, NADH4=1-504
GTR+G+FO, tRNA=505-656
GTR+G+FO, NADH5=657-898
```

The RAxML-NG partition file format is similar to that of standard RAxML and ExaML. Each line defines a partition, and contains the evolutionary model specification, the partition name, and the MSA site range(s).

IMPORTANT NOTE: the evolutionary model specification is not compatible with that in RAxML/ExaML! In particular, rate heterogeneity has to be defined for each partition individually, that is, we specify `GTR+G` for every partition with the Γ model instead of using a global `-m GTRGAMMA` switch on the command line as in standard RAxML/ExaML. Therefore, special care has to be taken when using legacy partition files.

Below, we show a more sophisticated example, where we use different per-partition substitution matrices and rate heterogeneity models, and also split the first gene by codon position:

```
$ cat prim2.part

GTR+G+FO, NADH4=1-504/3,2-504/3
JC+I, tRNA=505-656
GTR+R4+FC, NADH5=657-898
HKY, NADH4p3=3-504/3
```

Here, we use the *stride* notation to separate codon positions. For instance, `1-504/3` means *"every 3rd position in the range between 1 and 504'.*

6.2 Likelihood evaluation with partitioned models

Now, let us try to evaluate the likelihood on a fixed tree topology as in Section 5, but using a partitioned model (we will also increase the log output verbosity to be able to inspect the estimated parameter values):

```
$ raxml-ng --evaluate --msa prim.phy --threads 2 --model prim.part --tree
    T3.raxml.bestTree --prefix P1 -log verbose

Optimized model parameters:

   Partition 0: NADH4
   Speed (ML): 1.045481
   Rate heterogeneity: GAMMA (4 cats, mean),  alpha: 0.320532 (ML),
    weights&rates: (0.250000,0.007108) (0.250000,0.120533)
    (0.250000,0.628725) (0.250000,3.243634)
   Base frequencies (ML): 0.347608 0.343620 0.074289 0.234483
   Substitution rates (ML): 1.110014 16.895228 0.903118 0.001000 11.861976
     1.000000
```

```
   Partition 1: tRNA
   Speed (ML): 0.505287
   Rate heterogeneity: GAMMA (4 cats, mean),  alpha: 0.300774 (ML),
    weights&rates: (0.250000,0.005358) (0.250000,0.105097)
    (0.250000,0.597100) (0.250000,3.292444)
   Base frequencies (ML): 0.362527 0.230093 0.151307 0.256073
   Substitution rates (ML): 66.393654 308.024274 43.477166 37.411363
    671.608883 1.000000

   Partition 2: NADH5
   Speed (ML): 1.216009
   Rate heterogeneity: GAMMA (4 cats, mean),  alpha: 0.614255 (ML),
    weights&rates: (0.250000,0.056061) (0.250000,0.320104)
    (0.250000,0.888421) (0.250000,2.735414)
   Base frequencies (ML): 0.360963 0.322304 0.061324 0.255409
   Substitution rates (ML): 67.157660 1000.000000 56.903929 148.358484
    530.324413 1.000000
```

As we can see from the output above, even though we assigned the `GTR+G+FO` model to all three partitions, each of them has independent estimates of the parameter values (α shape parameter of the `GAMMA` distribution, base frequencies, and `GTR` substitution rates).

Let us repeat this evaluation using the second, more complex partition scheme:

```
$ raxml-ng --evaluate --msa prim.phy --threads 2 --model prim2.part --tree
     T3.raxml.bestTree --prefix P2 -log verbose
```

and compare the likelihoods for `P2` vs. `P1` vs. single `GTR+G+FO` model:

```
$ grep logLikelihood {E5,P1,P2}.raxml.log

E5.raxml.log:[00:00:00] Tree #1, final logLikelihood: -5709.002997
P1.raxml.log:[00:00:00] Tree #1, final logLikelihood: -5673.027260
P2.raxml.log:[00:00:00] Tree #1, final logLikelihood: -5673.868809
```

So `P1` has the best likelihood score, closely followed by `P2`. But both `P1` and `P2` also introduce more free parameters compared to `GTR+G+FO`:

```
$ grep "Free parameters" {E5,P1,P2}.raxml.log

E5.raxml.log:Free parameters (model + branch lengths): 30
P1.raxml.log:Free parameters (model + branch lengths): 50
P2.raxml.log:Free parameters (model + branch lengths): 52
```

Hence, we will once again use `AIC/BIC` criteria to assess the model complexity versus likelihood score trade-off (the lower the better):

```
grep "AIC score" {E5,P1,P2}.raxml.log

E5.raxml.log:AIC score: 11478.005995 / AICc score: 11480.151323 / BIC
    score: 11622.011097
P1.raxml.log:AIC score: 11446.054521 / AICc score: 11452.075772 / BIC
    score: 11686.063024
```

```
P2.raxml.log:AIC score: 11451.737617 / AICc score: 11458.260694 / BIC
    score: 11701.346461
```

The situation is less clear now: `AIC` and `AICc` favor the `P1` model, whereas `GTR+G+FO` has the best `BIC` score. Unfortunately, there seems to be no general consensus with respect to which information criterion is superior. Therefore, the decision whether to use `AIC`, `AICc` or `BIC` is left to the user. Furthermore, the computation of `AICc` and `BIC` scores requires the knowledge of *sample size*. In the context of phylogenetics, both the number of alignment sites (columns) and the total number of alignment characters (sites × taxa) have been proposed as sample size definitions (see e.g. Posada and Buckley, 2004, and references therein). In RAxML-NG, we define sample size as number of alignment sites, which is a more conservative option, also used by e.g., ModelTest-NG (Darriba et al., 2020) and IQTree (Nguyen et al., 2015).

6.3 Branch length linkage

In the output shown in Section 6.2, there is an extra parameter called `Speed` estimated for each partition:

```
Optimized model parameters:

   Partition 0: NADH4
   Speed (ML): 1.045481
[..]
   Partition 1: tRNA
   Speed (ML): 0.505287
[..]
   Partition 2: NADH5
   Speed (ML): 1.216009
```

What does it mean? In partitioned analyses, there are three common ways to estimate branch lengths (sometimes called *branch linkage* models):

- **linked**: all partitions share a common set of (global) branch lengths. This is the most simple model with the lowest number of parameters (*#branches*). However, it is often considered too unrealistic, as it is known that genes (or genome regions) evolve at different speeds.
- **unlinked**: each partition has its own, independent set of branch lengths. This model allows for the highest flexibility, but it also introduces a huge number of free parameters (*#branches* × *#partitions*), which makes it prone to overfitting.
- **scaled (proportional)**: a global set of branch lengths is estimated as in the "linked" mode, but each partition has an individual scaling factor; per-partition branch lengths are obtained by multiplying the global branch lengths with these individual scalers. This approach represents a compromise that allows to model distinct evolutionary rates across partitions while, at the same time, only introducing a moderate number of free parameters (*#branches* + *#partitions*).

RAxML-NG supports all three branch linkage models described above; they can be selected using the `--brlen` option. A recent simulation study by Duchene et al. (2018) showed that the `scaled` branch linkage model offers the best fit for a large number of typical representative datasets. This confirms the intuition about its "good" flexibility

versus complexity trade-off. Hence, RAxML-NG uses the `scaled` branch linkage model for partitioned analyses by default.

IMPORTANT NOTE: standard RAxML and ExaML use the linked branch length model by default. This should be kept in mind when comparing likelihoods and resulting topologies with those obtained via RAxML-NG!

So let us now explore how the `linked` and `unlinked` models behave on our toy dataset:

```
$ raxml-ng --evaluate --msa prim.phy --threads 2 --model prim.part --tree
    T3.raxml.bestTree --prefix P3 --brlen linked

$ raxml-ng --evaluate --msa prim.phy --threads 2 --model prim.part --tree
    T3.raxml.bestTree --prefix P4 --brlen unlinked
```

As could be expected, more complex models yield better likelihood scores (`unlinked` > `scaled` > `linked`):

```
$ grep logLikelihood {P1,P3,P4}.raxml.log

P1.raxml.log:[00:00:00] Tree #1, final logLikelihood: -5673.027260 <-
    scaled
P3.raxml.log:[00:00:00] Tree #1, final logLikelihood: -5678.429054 <-
    linked
P4.raxml.log:[00:00:00] Tree #1, final logLikelihood: -5648.348677 <-
    unlinked
```

However, the induced likelihood score difference is not always large enough to justify using additional model parameters:

```
grep "AIC score" {P1,P3,P4}.raxml.log

P1.raxml.log:AIC score: 11446.054521 / AICc score: 11452.075772 / BIC
    score: 11686.063024  <- scaled
P3.raxml.log:AIC score: 11452.858107 / AICc score: 11458.398743 / BIC
    score: 11683.266270  <- linked
P4.raxml.log:AIC score: 11476.697354 / AICc score: 11496.994752 / BIC
    score: 11908.712661  <- unlinked
```

Once again, we observe a disagreement between the `AIC/AICc` and `BIC` criteria, which choose `scaled` and `linked` branch length models, respectively. However, all three criteria exclude the extremely parameter-rich `unlinked` model.

6.4 Tree searches with partitioned models

In the previous subsection, we used partitioned models to re-evaluate the likelihood of the ML tree obtained under the `GTR+G` model. But what if we re-run tree search from scratch under a partitioned model? Will this alter the resulting likelihoods and/or topologies?

```
$ raxml-ng --msa prim.phy --model prim.part --prefix P5 --threads 2 --seed
     2 --brlen scaled

$ raxml-ng --msa prim.phy --model prim.part --prefix P6 --threads 2 --seed
     2 --brlen linked
```

```
$ raxml-ng --msa prim.phy --model prim.part --prefix P7 --threads 2 --seed
    2 --brlen unlinked
```

Checking the new likelihood scores

```
$ grep "Final LogLikelihood" {P5,P6,P7}.raxml.log

P5.raxml.log:Final LogLikelihood: -5672.951995
P6.raxml.log:Final LogLikelihood: -5678.301081
P7.raxml.log:Final LogLikelihood: -5648.204296
```

shows that they are almost identical to the values obtained on the `T3` topology (see Section 6.2). Moreover, all three partitioned runs converged to the same ML tree topology as the unpartitioned `T3` run (see Section 3).

Of course, this observation will not hold for all datasets. However, there is some evidence that the choice of the DNA substitution models has a rather limited influence on the resulting tree topology (Hoff et al., 2016). Of course, this result does not mean that model selection should not be conducted. However, it suggests that subtle details such as the conflicts between `AIC(c)` versus `BIC` or sample size definition are of minor concern in practice.

7 Parallelization and performance

7.1 Introduction

RAxML-NG supports three levels of parallelism: CPU instruction level (vectorization), intra-node (multithreading), and inter-node (MPI) parallelism. Unlike standard RAxML/ExaML, a single RAxML-NG executable offers *all* parallelism levels. The desired parallelism level can be configured at run-time (MPI support is optional and should be enabled at compile-time).

As of v.0.8.0*b*, RAxML-NG only supports *fine-grained* parallelization across MSA sites. This is the same parallelization approach that has been used in the PThreads version of standard RAxML and ExaML. It is conceptually different from the *coarse-grained* parallelizations across independent tree searches or tree moves as implemented in RAxML-MPI or IQTree-MPI (Nguyen et al., 2015), respectively. With fine-grained parallelization, the number of CPU cores that can be efficiently utilized is limited by the MSA "width" (=number of site patterns). For instance, using 20 cores on a single-gene protein alignment with 300 sites would be suboptimal, and using 100 cores would most probably result in a huge slowdown. In order to prevent wasting CPU time and energy, RAxML-NG will warn you – or, in extreme cases, even refuse to run – if you try to assign too few MSA sites to a core.

Coarse-grained parallelization, although not directly implemented in RAxML-NG, can be easily emulated as shown in Section 7.7.

7.2 Multithreading (pthreads)

By default, RAxML-NG will start as many threads as there are CPU cores available on your system. Most modern CPUs employ so-called *hyperthreading* technology, which makes each *physical* core appear as two *logical* cores to software. Hyperthreading can be beneficial for some programs, but **RAxML-NG achieves the best performance when run with one thread per physical core**. Therefore, RAxML-NG will try to detect if CPU supports hyperthreading, and will reduce the number of threads accordingly.

For instance, on a laptop with an Intel i7-8550U processor, RAxML-NG will detect 4 (physical) cores and use 4 threads by default:

```
  parallelization: PTHREADS (4 threads), thread pinning: OFF
```

even though this CPU has 8 logical cores:

```
$ lscpu -e

CPU NODE SOCKET CORE L1d:L1i:L2:L3 ONLINE MAXMHZ    MINMHZ
0   0    0      0    0:0:0:0       yes    4000,0000 400,0000
1   0    0      1    1:1:1:0       yes    4000,0000 400,0000
2   0    0      2    2:2:2:0       yes    4000,0000 400,0000
3   0    0      3    3:3:3:0       yes    4000,0000 400,0000
4   0    0      0    0:0:0:0       yes    4000,0000 400,0000
5   0    0      1    1:1:1:0       yes    4000,0000 400,0000
6   0    0      2    2:2:2:0       yes    4000,0000 400,0000
7   0    0      3    3:3:3:0       yes    4000,0000 400,0000
```

Unfortunately, it is very hard to reliably detect situations when hyperthreading is *supported* by the CPU, but *disabled* in BIOS. For instance, this setup can be found on Amazon AWS as well as on some clusters. In this situation, RAxML-NG can underestimate the number of available physical cores. Thus, it is recommended to use the `--threads` option and manually set the number of threads, to be on the safe side.

7.3 MPI and hybrid MPI/pthreads

If compiled with MPI support, RAxML-NG can leverage multiple compute nodes for a single analysis. Please check your cluster documentation for system-specific instructions on running MPI programs as this is different for every cluster. In MPI-only mode, you should start one MPI process *per physical CPU core* (the number of threads will be set to one by default).

However, in most cases, a hybrid MPI/pthreads setup will be more efficient in terms of both, runtime, and memory consumption. Typically, you would start one MPI rank *per compute node*, and one thread per physical core (e.g. `--threads 16` for nodes equipped with dual-slot octa-core CPUs). Below is a sample job submission script for the cluster at our research institute which uses SLURM (Yoo et al., 2003) as workload manager:

```
#!/bin/bash
#SBATCH -N 4
#SBATCH -B 2:8:1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=16
#SBATCH --hint=compute_bound
#SBATCH -t 08:00:00

raxml-ng-mpi --msa rbcl.phy --model GTR+G --prefix H1 --threads 16
```

Here, we requested 4 nodes with one task per node and 16 cores per node/task. We further specify job runtime limit (`-t`), set thread mapping according to node topology (`-B`) and prevent core oversubscription with `--hint=compute_bound` (please refer to https://slurm.schedmd.com/documentation.html for the full list of SLURM options).

Once again, please consult your cluster documentation to find out how to properly configure a hybrid MPI/pthreads run. **Please note that incorrect configuration can result in extreme slowdowns and hence waste time and resources!**

7.4 Thread pinning

For attaining optimal performance, it is crucial to ensure that only one RAxML-NG thread is running on each physical CPU core. Usually, the operating system can handle the thread-to-core assignment automatically. However, some (misconfigured) MPI runtimes tend to pack all threads onto a single CPU core, resulting in abysmal performance. To avoid this situation, each thread can be "pinned" (explicitly assigned to) to a particular CPU core.

In RAxML-NG, thread pinning is enabled by default only in the hybrid MPI/pthreads mode when one MPI rank per node is used. You can explicitly enable or disable thread pinning with `--extra thread-pin` and `--extra thread-nopin`, respectively.

7.5 Vector instructions

RAxML-NG will automatically detect the best (fastest) set of vector instructions available on your CPU, and use the respective computational kernels to achieve optimal performance. On modern Intel CPUs, this autodetection mechanism appears to work pretty well, so most probably you will not need to worry about this. However, you can force RAxML-NG to use a specific set of vector instructions with the `--simd` option, for instance,

```
$ raxml-ng --msa prim.phy --model GTR+G --prefix V1 --threads 2 --seed 2
    --simd sse
```

to use SSE3 vector instructions, or

```
$ raxml-ng --msa prim.phy --model GTR+G --prefix V2 --threads 2 --seed 2
    --simd none
```

to use non-vectorized (scalar) instructions. This option might be useful for debugging, but otherwise using non-optimal vectorization should be avoided as it incurs a substantial performance penalty:

```
$ grep "Elapsed time:" {T3,V1,V2}.raxml.log

T3.raxml.log:Elapsed time: 7.802 seconds  <- AVX (autodetect)
V1.raxml.log:Elapsed time: 15.394 seconds <- SSE
V2.raxml.log:Elapsed time: 21.663 seconds <- scalar
```

7.6 Determining the optimal number of threads

One of the most frequent question we get from RAxML users is: *How many threads should I use?*. As always, simple questions are the toughest ones. You might as well ask: *How fast should I drive?*. In both cases, the answer would be: *It depends*. It depends on where you drive (dataset), your vehicle (system), and your priorities (time versus money/energy). In RAxML-NG, we have implemented some warning signs and radar speed guns, for your own safety. As in real life, you are free to ignore them (with the `--force` option), which can result in two things: (1) earlier arrival, or (2) lost time and money. Fortunately, unlike on the road, you can experiment with RAxML-NG safely, and we encourage you to do so.

1.3:20 RAxML-NG

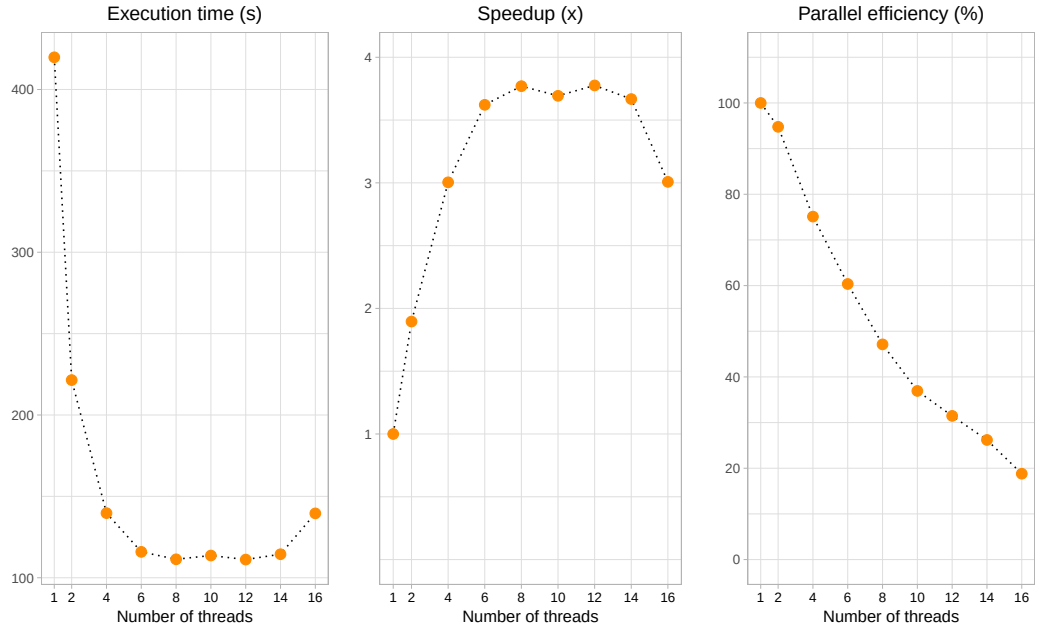


Figure 1 Typical scaling of RAxML-NG on a small alignment (here: 436 taxa and 1,371 DNA sites). Parallel efficiency (*right*) shows percentage of ideal (linear) speedup and is defined as follows: $Parallel\_efficiency = Speedup/Number\_of\_threads \times 100\%$.

A reasonable workflow for analyzing a large dataset would be as follows. First, run RAxML-NG in parser mode, that is,

```
$ raxml-ng --parse --msa rbcl.phy --model GTR+G+F --prefix rbcl
```

This command will generate a binary MSA file (`rbcl.raxml.rba`), which can subsequently be loaded by RAxML-NG much faster than the original FASTA alignment. Furthermore, it will print the estimated memory requirements and the recommended number of threads for this dataset:

```
[00:00:00] Reading alignment from file: rbcl.phy
[00:00:00] Loaded alignment with 436 taxa and 1371 sites

Alignment comprises 1 partitions and 1001 patterns

NOTE: Binary MSA file created: rbcl.raxml.rba

* Estimated memory requirements                : 54 MB
* Recommended number of threads / MPI processes: 4
```

The recommended number of threads is computed using a simple heuristic and should yield a decent runtime/resource utilization trade-off in most cases. It also constitutes a good starting point for your experiments: you can now run tree searches with a varying number of threads, for instance,

```
$ raxml-ng --search1 --msa rbcl.raxml.rba --seed 1 --threads 2
$ raxml-ng --search1 --msa rbcl.raxml.rba --seed 1 --threads 4
```

```
$ raxml-ng --search1 --msa rbcl.raxml.rba --seed 1 --threads 8
```

and so on. For a small example dataset as in our example, the execution time will decrease initially as we add more threads, but will then quickly level off (leftmost plot below). Although the maximum speedup of $\approx 3.75\times$ can be attained with $8-14$ threads (middle plot), it induces a rather poor parallel efficiency of 60%-30% (right plot; parallel efficiency is defined as speedup divided by the number of threads). The recommended number of threads (4) yields a reasonable speedup ($3\times$) without compromising the parallel efficiency too much (75%). Finally, if we use an excessively large number of cores (≥ 16 in this example), execution times will start to *increase* again. Although the actual speedups will vary across datasets and systems, the general trend will stay the same. Therefore, it is up to the user to decide how many resources (=higher CPU time) can be sacrificed to obtain the results faster (=lower execution time).

7.7 Coarse-grained parallelization for short alignments

If you want to utilize a large number of CPU cores for analyzing a small ("single-gene") alignment, please have a look at our ParGenes pipeline (Morel et al., 2019) which implements coarse-grained parallelization and dynamic load balancing. ParGenes is freely available at https://github.com/BenoitMorel/ParGenes.

Alternatively, coarse-grain parallelization can easily be emulated by executing multiple RAxML-NG instances, but with distinct random seeds. For instance, let us assume that we want to run an "all-in-one' analysis on the dataset described in Section 7.6, and we want to use a server with 16 CPU cores. As Figure 1 shows, the fine-grained parallelization across 16 cores is very inefficient for this dataset. We will therefore use fine-grained parallelization with 2 cores per tree search, which means we can run $16/2 = 8$ RAxML-NG instances in parallel. First, we will infer 24 ML trees, using 12 random and 12 parsimony-based starting trees. Hence, each RAxM-NG instance will run searches from $24/8 = 3$ starting trees. Below is a sample `SLURM` script for doing this:

```
#!/bin/bash
#SBATCH -N 1
#SBATCH -n 8
#SBATCH -B 2:8:1
#SBATCH --threads-per-core=1
#SBATCH --cpus-per-task=2
#SBATCH -t 02:00:00

for i in `seq 1 4`;
do
  srun -N 1 -n 1 --exclusive raxml-ng --search --msa rbcl.raxml.rba --tree
     pars{3}    --prefix CT$i --seed $RANDOM --threads 2 &
done

for i in `seq 5 8`;
do
  srun -N 1 -n 1 --exclusive raxml-ng --search --msa rbcl.raxml.rba --tree
     rand{3}   --prefix CT$i --seed $RANDOM --threads 2 &
done
```

```
wait
```

Of course, this script has to be adapted for your specific cluster configuration and/or job submission system. You can also use `GNU parallel`, or directly start multiple RAxML-NG instances from the command line. Please pay attention to the ampersand symbol (`&`) at the end of each RAxML-NG command line: it is extremely important here, since if you forget the ampersand all RAxML-NG instances will run one after another and *not* in parallel! Furthermore, we add `--exclusive` flag to tell ensure that raxml-ng instances will be assigned to *distinct* CPU cores (this is default behavior with some SLURM configurations, but not always).

Once the job has finished, we can inspect the likelihoods:

```
$ grep "Final LogLikelihood" CT*.raxml.log | sort -k 3

CT7.raxml.log:Final LogLikelihood: -30621.004116
CT6.raxml.log:Final LogLikelihood: -30621.537107
CT2.raxml.log:Final LogLikelihood: -30621.699234
CT3.raxml.log:Final LogLikelihood: -30622.534482
CT1.raxml.log:Final LogLikelihood: -30622.783250
CT8.raxml.log:Final LogLikelihood: -30623.963471
CT5.raxml.log:Final LogLikelihood: -30623.020351
CT4.raxml.log:Final LogLikelihood: -30623.378857
```

and select the best-scoring tree (`CT7.raxml.bestTree` in our case):

```
$ ln -s CT7.raxml.bestTree best.tre
```

The same trick can be applied to bootstrapping. For the sake of simplicity, let us infer $8 \times 15 = 120$ replicate trees:

```
for i in ‘seq 1 8‘;
do
  raxml-ng --bootstrap --msa rbcl.raxml.rba --bs-trees 15 --prefix CB$i --
    seed $RANDOM --threads 2 &
done

wait
```

Now, we can simply concatenate all replicate tree files (`*.raxml.bootstraps`) and then proceed with the bootstrap convergence check as well as branch support calculation as usual (see Section 4):

```
$ cat CB*.raxml.bootstraps > allbootstraps

$ raxml-ng --bsconverge --bs-trees allbootstraps --prefix CS --seed 2 --
    threads 1

$ raxml-ng --support --tree best.tre --bs-trees allbootstraps --prefix CS
    --threads 1
```

There are two things to keep in mind when conducting this type of coarse-grained parallelization. First, memory consumption will grow proportionally to the number of

RAxML-NG instances running in parallel. That is, in our case, an estimate given by the `--parse` command should be multiplied by 8. Second, correct thread allocation (one thread per CPU core) is crucial for achieving the optimal performance. Hence, we recommend to check thread allocation, for instance, by running `htop` after your initial script submission.

NOTE: RAxML-NG v. 1.0 and later provides "built-in" coarse-grained parallelization, please consult online documentation for details: https://github.com/amkozlov/raxml-ng/wiki/Parallelization.

Acknowledgements

The authors wish to thank the following former students of our 2018 summer school on computational molecular evolution for useful comments on the initial draft of this book chapter: Sunitha Manjari and Loïc Meunier.

References

Czech, L., Huerta-Cepas, J., and Stamatakis, A. (2017). A critical review on the use of support values in tree viewers and bioinformatics toolkits. *Molecular Biology and Evolution*, 34(6):1535–1542.

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, 37(1):291–294.

Duchene, D. A., Tong, K. J., Foster, C. S., Duchene, S., Lanfear, R., and Ho, S. Y. (2018). Linking branch lengths across loci provides the best fit for phylogenetic inference. *bioRxiv*.

Hoff, M., Orf, S., Riehm, B., Darriba, D., and Stamatakis, A. (2016). Does the choice of nucleotide substitution models matter topologically? *BMC Bioinformatics*, 17(1):143.

Huson, D. H. and Scornavacca, C. (2012). Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biology*, 61(6):1061–1067.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y., Faircloth, B. C., Nabholz, B., Howard, J. T., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., and Jermiin, L. S. (2017). Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6):587.

Kozlov, A. M., Aberer, A. J., and Stamatakis, A. (2015). ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics*, 31(15):2577–2579.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Kozlov, O. (2018). *Models, Optimizations, and Tools for Large-Scale Phylogenetic Inference, Handling Sequence Uncertainty, and Taxonomic Validation*. PhD thesis, Karlsruher Institut für Technologie (KIT).

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2016). Partitionfinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, 34(3):772–773.

Lemoine, F., Domelevo Entfellner, J. B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., and Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, 556(7702):452–456.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763–767.

Morel, B., Kozlov, A. M., and Stamatakis, A. (2019). ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics*, 35(10):1771–1773.

Nguyen, L.-T. et al. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M., and Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of Computational Biology*, 17(3):337–354. PMID: 20377449.

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808.

Rambaut, A. (2012). Figtree v1. 4. molecular evolution, phylogenetics and epidemiology. *Edinburgh, UK: Retrieved from http://tree. bio. ed. ac. uk/software/figtree [Google Scholar]*.

Robinson, D. and Foulds, L. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131 – 147.

Shimodaira, H. and Hasegawa, M. (2001). Consel: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17(12):1246–1247.

Stamatakis, A. (2006a). Phylogenetic Models of Rate Heterogeneity: A High Performance Computing Perspective. In *Proc. of IPDPS2006*, HICOMB Workshop, Proceedings on CD, Rhodos, Greece.

Stamatakis, A. (2006b). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stamatakis, A. and Aberer, A. (2013). Novel parallelization schemes for large-scale likelihood-based phylogenetic inference. In *Parallel Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on*, pages 1195–1204.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics*, 139(2):993–1005.

Yoo, A. B., Jette, M. A., and Grondona, M. (2003). Slurm: Simple linux utility for resource management. In *Workshop on Job Scheduling Strategies for Parallel Processing*, pages 44–60. Springer.

A Summary of changes compared to RAxML 8.x

RAxML-NG offers multiple improvements and extensions compared to standard RAxML 8.x. On the other hand, not all features of standard RAxML are implemented as of RAxML-NG v0.8.0b (most notably, rapid bootstrapping and CAT/PSR model [Stamatakis 2006a]). Furthermore, several important defaults have been changed to be in line with best practices: for instance, RAxML-NG is using multiple starting trees and scaled branch lengths by default. To give a better overview, we provide a side-by-side comparison of standard RAxML 8.x and RAxML-NG 0.8.0b in Table 1.

Option/Feature	RAxML 8.x	RAxML-NG 0.8.0b
Features		
Bootstrapping	standard, rapid	standard
Parallelization scheme	fine-grained (PTHREADS) coarse-grained (MPI)	fine-grained (PTHREADS and MPI)
Checkpointing	NO	YES
Binary MSA	NO	YES
Evolutionary models		
Rate heterogeneity across sites (RHAS)	GAMMA, p-inv, CAT	GAMMA, p-inv, FreeRate
RHAS linkage	global	per-partition
Branch length linkage	linked, unlinked	linked, unlinked, scaled
LG4X with linked branches	YES	NO
User-specified parameter values	NO	YES
Defaults		
Starting tree(s)	parsimony (1)	parsimony(10)+random(10)
Stationary state frequencies	empirical	ML estimate
Branch lengths linkage	linked	scaled
Number of boostrap replicates	100	AUTO (bootstopping)

Table 1 Differences in features and default settings between standard RAxML and RAxML-NG v0.8.0b

Chapter 1.4 The Bayesian Approach to Molecular Phylogeny

Nicolas Lartillot
Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive,
43 Bld du 11 Novembre 1918, 69622 Villeurbanne cedex, France.
nicolas.lartillot@univ-lyon1.fr
https://orcid.org/0000-0002-9973-7760

Abstract

Bayesian inference is now routinely used in phylogenomics and, more generally, in macroevolutionary studies. Beyond the philosophical debates it has raised concerning the choice of the prior and the meaning of posterior probabilities, Bayesian inference, combined with generic Monte Carlo algorithms, offers a flexible framework for introducing subjective or context information through the prior, but also, for designing hierarchical models formalizing complex patterns of variation (across sites or branches) or the integration of multiple levels of evolutionary processes. In this chapter, the principles of Bayesian inference, such as applied to phylogenetic reconstruction, are first introduced, with an emphasis on the key features of the Bayesian paradigm that explain its flexibility in terms of model design and its robustness in inferring complex patterns and processes. A more specific focus is then put on the question of modeling pattern-heterogeneity across sites, using both parametric and non-parametric random-effect models. Finally, the current computational challenges are discussed.

How to cite: Nicolas Lartillot (2020). The Bayesian Approach to Molecular Phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 1.4, pp. 1.4:1–1.4:17. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Introduction

Bayesian inference was introduced in phylogenetics in the mid 90's (Yang and Rannala, 1997; Mau et al., 1999; Larget and Simon, 1999; Li et al., 2000; Huelsenbeck and Ronquist, 2001). From the very beginnings, it has motivated a lot of discussion about its merits and drawbacks, compared to more firmly established approaches such as maximum likelihood (reviewed by Holder and Lewis (2003); Yang (2007) and in Chapter 1.2 [Stamatakis and Kozlov 2020]). Part of the discussions then revolved around philosophical or foundational issues: how to choose the priors and how to have a control on the sensitivity of the analysis to this choice? What is the meaning of posterior probabilities? How do those compare with alternative measures of statistical support, like non-parametric bootstrap?

Meanwhile, Bayesian inference has reached the stage of practical applications over a broad array of research questions in phylogenomics and in evolutionary studies. To this end, much work has been devoted to the design of increasingly sophisticated models and to the development of efficient algorithms based on Markov Chain Monte Carlo (MCMC). As a result, Bayesian inference, and its relevance to evolutionary genetics, can now be better understood based on its practical impact on current research in our field. As it turns out, the use of Bayesian inference has substantially renewed our perspective on the role of models in phylogenomics (see Chapter 2.1 [Simion et al. 2020]), offering new opportunities that were not directly reachable using classical approaches. Conversely, these practical applications

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 1.4; pp. 1.4:1–1.4:17

A book completely handled by researchers.
No publisher has been paid.

have shown some weaknesses and limitations of the purely Bayesian philosophical stance, while emphasizing its connections and its similarity with the maximum likelihood paradigm.

The aim of the present chapter is, first, to briefly introduce the basics of Bayesian inference, such as applied to phylogenetics, and to point out its main features in this specific context. In a second step, we will focus on one particular application of Bayesian inference in phylogenomics, namely, the development of Bayesian non-parametric models accounting for variation across sites in amino acid preferences. Finally, we will discuss the current challenges and propose some perspectives concerning how these challenges are being tackled by current statistical and computational research.

2 Bayesian inference in phylogenetics

2.1 General principles of probabilistic inference

Consider a simple phylogenetic problem, in which the aim is to infer the phylogeny of a group of taxa, based on a multiple sequence alignment for a single gene (e.g. ribosomal RNA), and assuming a simple one-parameter process of sequence evolution, the Kimura model (which depends only on the transition-tranversion rate ratio, see Chapter 1.1 [Pupko and Mayrose 2020]). Let us denote by D the sequence alignment, T the unknown tree topology, with branch lengths noted l, and κ the transition/transversion rate ratio. The likelihood is defined as the probability that the sequence evolutionary process with transition-transversion rate ratio κ, running along tree T with branch lengths l, produces the nucleotide sequences D at the tips of the tree. This probability can be written:

$$L(T, l, \kappa) \quad = \quad p(D \mid T, l, \kappa).$$

In the present case, sites are assumed to evolve independently of each other. As a result, the likelihood can be written as a product over all sites. If D_i stands for the site pattern observed at position i, for $i = 1 \ldots n$, where n is the number of aligned positions, then:

$$p(D \mid T, l, \kappa) = \prod_{i=1}^{n} p(D_i \mid T, l, \kappa).$$

The maximum likelihood approach to tree estimation works as follows. First, for a given tree T, the likelihood is maximized with respect to the branch lengths and the parameters of the substitution process:

$$\hat{L}(T) \quad = \quad \max_{l,\kappa} L(T, l, \kappa).$$

This defines a likelihood score for a given tree topology T. Then, we search for the tree topology T which maximizes this score, i.e. the aim is to find the tree $\hat{T}$ such that:

$$\hat{L}(\hat{T}) \quad = \quad \max_{T} \hat{L}(T).$$

Of note, the likelihood is jointly maximized with respect to all unknowns, those we are interested in (here, the tree topology T) and those that are not of direct interest but have an influence on the probability of producing the data (nuisance parameters, here, l and κ). As a result, a tree topology is scored based only on the best-case scenario for all unknown aspects of the evolutionary process under this tree topology, without any consideration for alternative configurations of these unknown nuisances. This is an important point that distinguishes maximum likelihood and Bayesian inference, as we will now see.

2.2 The Bayesian approach

How does Bayesian inference proceed in the present case? First, one has to define a prior distribution over the tree topologies, the branch lengths and the parameter of the substitution model. These priors can be denoted as $p(T)$, $p(l)$ and $p(\kappa)$. Of note, $p(T)$ is a probability over the discrete space of tree topologies. On the other hand, since l and θ are continuous parameters, $p(l)$ and $p(\theta)$ are not probabilities, but probability densities. Which priors have more specifically been used in practical applications will be discussed below. In the present case, one would typically assume simple priors like the following: a uniform prior over T, an exponential of mean 0.1 for branch lengths, and a uniform prior between 0 and 20 for the transition-transversion rate ratio κ.

Second, for a given tree topology T, the likelihood is averaged over all possible values for the continuous parameters l and κ, weighted by the prior distributions over these two parameter components. This defines the marginal likelihood of tree topology T:

$$p(D \mid T) = \int_\theta \int_l p(D \mid T, l, \kappa) p(l) p(\kappa) dl d\kappa. \quad (1)$$

Finally, using Bayes theorem, we can define the posterior probability of the tree topology T:

$$p(T \mid D) = \frac{p(D \mid T)\, p(T)}{p(D)},$$

where the denominator $P(D)$ is the marginal probability of the data (or marginal likelihood), obtained by summing the numerator over all possibles tree topologies (so as to normalize the posterior probability distribution):

$$p(D) = \sum_T p(D \mid T)\, p(T).$$

In many situations, the normalization factor $p(D)$ is not essential, and a simpler account of Bayes theorem is then given by:

$$p(T \mid D) \propto p(D \mid T)\, p(T), \quad (2)$$

which essentially states that the posterior probability of a tree T is proportional to its prior probability $p(T)$ multiplied by the weight of evidence contributed by the data, $p(D \mid T)$. Thus, if $p(T)$ represents our state of belief about which trees are likely to be correct before we have seen the data, then Bayes theorem formalizes how one would update our state of belief upon seeing data D, leading to our posterior state of belief $p(T \mid D)$. For very large datasets, the posterior will typically be concentrated on one single tree topology. In terms of beliefs, this amounts to a nearly complete certainty about that tree being the correct phylogeny, given the data and the model. For smaller datasets, on the other hand, multiple trees may each receive a significant proportion of the total posterior probability mass, which then represents our remaining uncertainty about the phylogenetic history of our clade of interest, given the data.

The equations above could be equivalently rewritten, first by defining a joint posterior over the tree topology and the continuous parameters, using Bayes theorem:

$$p(T, l, \theta \mid D) \propto p(D \mid T, l, \kappa)\, p(T)\, p(l)\, p(\kappa). \quad (3)$$

This formulation is more general, as it puts all unknowns, tree topology, branch lengths, parameters of the sequence evolutionary process, on the same footing. Marginalization is done only in a second step, in a way that depends on the question being asked. Thus far, we

have assumed that the topology was of interest, and thus, we have focussed on the marginal posterior on T, which is obtained by integrating out l and κ:

$$p(T \mid D) = \int_{\kappa} \int_{l} p(T, l, \kappa \mid D) dl d\kappa. \quad (4)$$

Of note, this definition of the marginal posterior on T is equivalent to that given by Eq. 2 above. If instead we were interested in estimating the transition/transversion rate ratio κ, then we would consider the marginal posterior distribution over κ, which is summed over all possible tree topologies and branch lengths:

$$p(\kappa \mid D) = \sum_{T} \int_{l} p(T, l, \kappa \mid D) dl. \quad (5)$$

In the end, Bayesian inference always reduces to computing the posterior probability of what we want to know, given the available evidence and the structural assumptions of the generating model, and this, integrated (averaged) over all other unknowns.

2.3 Practical Bayesian inference using Monte Carlo

The equations indicated above involve sums over a large number of alternative tree topologies and, for a given topology T, integrals over all possible branch lengths and all values for κ. In general, those integrals are not analytically available and would be difficult to numerically evaluate with sufficient accuracy. As a practical alternative, Bayesian inference is most often implemented using Monte Carlo approaches.

The general aim of Monte Carlo is to design random sampling algorithms targeting a probability distribution of interest – here, the joint posterior distribution (Eq. 3). By far the most commonly used algorithm is the Markov chain Monte Carlo (MCMC) approach. The idea of MCMC is to implement a random walk in the parameter space of the model, such that parameter configurations are visited at a frequency proportional to their posterior probability. Running the algorithm for a sufficiently long time yields samples from the posterior distribution:

$$(T_j, l_j, \kappa_j) \sim p(T, l, \kappa \mid D)$$

for $j = 1 \ldots N$, with N a suitably large number of samples. In the case of MCMC, the samples are not independent (successive samples are typically correlated). Furthermore, they are from the targeted posterior distribution only asymptotically. In practice, this means that, starting from an arbitrary parameter configuration, the chain reaches its stationary state only after a burn-in period, and it is only after this stationary state has been reached that samples can be considered to be representative draws from the posterior.

Once such a large sample has been obtained, any marginal over the posterior probability can then be approximated simply by averaging over this sample. For instance, the frequency of a given tree topology T in the sample will be a Monte Carlo estimate of the posterior probability $p(T \mid D)$. More generally, the frequency at which a given group of taxa is found monophyletic in the sample is a Monte Carlo estimate of the posterior probability that this clade is monophyletic. Accordingly, a convenient way to summarize the analysis is to draw the majority-rule consensus of all trees collected by Monte Carlo and label each clade with the Monte Carlo estimate of its posterior probability support.

If, on the other hand, our interest is in some continuous parameters, say the transition/transversion rate ratio, then the histogram of the values of κ collected in the Monte Carlo represents our estimate of the posterior probability density over this parameter. From

there, a 95% credible interval can be computed, by sorting the samples by increasing value and excluding the 2.5% most extreme values at both ends.

2.4 Some important properties of Bayesian inference

Based on the general description of Bayesian inference given above, several points are worth pointing out and discussing.

Averaging over uncertainty

First, Bayesian inference, when conducted on some parameter of interest, always averages uncertainty over all other nuisance parameters. This has already been emphasized above (Eqs. 4 and 5), but some intuition can also be gained from how the output of the MCMC is processed. For instance, as was just pointed out, the Monte Carlo estimate of the posterior probability of a given clade is just the frequency at which this clade is present in the trees sampled by Monte Carlo. Importantly, all trees presenting this specific clade might differ from each other in many other respects – in terms of branch lengths, parameter values, but also in terms of the other clades that are present elsewhere in the tree. By this, we see that, in our evaluation of how likely it is that a given group is monophyletic, we have averaged our evaluation over many possible outcomes for all other aspects of the problem – in some sense, we have diversified our inference portfolio. This is in sharp contrast with maximum likelihood, which bets on one single configuration for the nuisance parameters when deciding for its point estimate. Averaging over uncertainty is expected to lead to more robust inference, in particular, when there are many nuisance parameters (Huelsenbeck et al., 2000).

Monte Carlo versus analytical integration

The point just discussed also shows a key relation between Monte Carlo and integration. Namely, the simple fact of sampling from the joint posterior over tree topology and other parameters and then discarding all parameters, keeping only the tree topology, is equivalent to sampling tree topologies from their marginal posterior distribution. In other words, Monte Carlo automatically implements Eq. 4, and this, without ever explicitly calculating this integral. This apparently anecdotal mathematical observation has important practical consequences: whenever a likelihood is a complex integral over many nuisance variables, then, instead of explicitly calculating this integral, it is always possible to explicitly sample from the nuisance variables, jointly with the parameter of interest. This approach of parameter expansion (or data augmentation) makes it possible to implement a broad category of models that would otherwise not be accessible by explicit numerical integration.

Priors

As mentioned above, averaging the likelihood over the nuisance parameters is expected to lead to more robust inference. On the other hand, this average is taken over a specific prior. More generally making decisions based on posterior probabilities, which are directly proportional to the prior, raises the question of how to choose the priors and how the inference will depend on this choice. Prior choice is perhaps the most important question in Bayesian inference, with many consequences, both conceptual and practical. There is a vast literature on the question, and many questions are still open. The problem is complex, since there are in fact very different approaches to prior definition, proceeding from different philosophies and resulting in posterior probabilities that do not have the same operational properties or the

same meaning. As a tentative typology, it may be useful to distinguish between the following approaches to prior elicitation:

- Informative priors based on expert knowledge. These priors are typically advocated by the so-called subjective Bayesian school, initiated by De Finetti (1974). These priors are not so often used in phylogenetics, with the important exception of soft fossil calibrations in molecular dating (see Yang and Rannala (2006); Chapter 5.1 [Pett and Heath 2020]).
- Uninformative, default or reference priors. The various names given to these priors express their slightly different objectives (priors expressing lack of information, meant to be used by default when expert knowledge is not available, or providing a neutral reference for summarizing empirical information), but in the end, all converge toward very similar practical recommendations. These priors define what is sometimes referred to as the objective Bayesian philosophy (Berger, 2006). In practice, many priors used in Bayesian phylogenetics are meant to be uninformative, or at least sufficiently vaguely informative, so that they can be used by default (without invoking case-dependent expertise or prior information). For instance, the most widely used prior for reconstructing phylogenies in an undated context is a uniform prior over all unrooted tree topologies (Huelsenbeck and Ronquist, 2001) .
- Hierarchical priors. These priors correspond to the closely related empirical or hierarchical Bayesian philosophies. A good example in phylogenetics is to allow for uncorrelated gamma distributed rates across branches in a relaxed clock analysis (Lepage et al., 2007). In this context, branch-specific substitution rates share the same gamma prior. In turn, the shape and scale parameters of this gamma prior, which tune the mean and the variance of rates across branches, are also unknown. Accordingly, a second-stage prior is invoked over these two hyper-parameters. This hyper-prior is typically chosen to be vaguely informative, and as a result, the posterior distribution on the shape and scale parameters will be mostly dictated by the signal about rate variation contained in the sequence data. Hierarchical priors thus represent a powerful tool for designing models allowing for complex modulations of the evolutionary process across branches, sites, or genes, in a way that will be automatically tuned to the true amount of variation present in the data.
- Mechanistic priors, i.e. priors that are themselves justified on the grounds of some macro-evolutionary mechanism or process. A good example is the birth-death prior over phylogenetic trees in a dated context (Yang and Rannala, 1997), which essentially implements a model of species diversification with constant speciation and extinction rates. In a sense, mechanistic priors proceed from the realization that our prior knowledge for some parameter of our problem (here, the unknown phylogeny) is best formalized in terms of a generating model – thus not unlike the likelihood itself. In the end, the result is not very different from the mixed models typically considered in a maximum likelihood context.

Flexibility in model design

The use of mechanistically inspired and hierarchical priors, combined with generic Monte Carlo approaches, has an important practical consequence. Indeed, it makes it possible to design hierarchical models, articulating together multiple levels of processes and integrating multiples sources of empirical data. In this direction, many important developments have taken place over the last decade, including the following:

- Brownian processes for relaxing the molecular clock (see Thorne and Kishino (2002); Chapter 4.4 [Bromham 2020]);

- birth-death models for describing speciation-extinction-fossilization processes (see Heath et al. (2014); Chapter 5.1 [Pett and Heath 2020]);
- integration of the comparative method (models of trait evolution) with the relaxed molecular clock (Lartillot and Poujol, 2011);
- integration of morphological and genetic sequence data (total-evidence dating) (Zhang et al., 2016)
- explicit models of gene duplication, loss and horizontal transfer over species trees (see Akerborg et al. (2009); Chapter 3.2 [Boussau and Scornavacca 2020]);
- multi-species coalescent approaches (see Yang and Rannala (2010); Heled and Drummond (2010); Chapter 3.3 [Rannala et al. 2020]);
- priors over viral phylogenies derived from epidemiological models (see (Kühnert et al., 2014); Chapter 5.3 [Zhukova et al. 2020]);
- non-parametric models for modeling random-effects across sites (Lartillot and Philippe, 2004; Huelsenbeck and Suchard, 2007) or branches (Heath et al., 2012).

The list is not exhaustive, and we can expect many new developments along similar lines. Some of these integrative or hierarchical models are introduced in other chapters of this book (see references above).

3 Bayesian non-parametric site-heterogeneous models

3.1 Variation across sites

One of the most prominent features of multiple sequence alignments at large evolutionary scale is the amount of variation across sites in the degree and the patterns of conservation. This can be seen both at the nucleotide and the amino acid levels. The reasons for this are well understood: sequences that are conserved over long evolutionary periods are almost certainly under strong purifying selection. Selection, however, is highly context-specific, and is thus likely to be fairly disparate across sites of a gene, both in overall intensity and in the nature of the preferred nucleotides or amino acids. This problem is further amplified by the fact that phylogenetic reconstruction is normally conducted using those genes and gene regions that can be reliably aligned over a broad phylogenetic scale. Selecting well-aligned sequences is essential, in order to guarantee the validity of the assumption that sequence variation in the data matrix is only caused by point substitutions, an assumption made by virtually all models currently used in phylogenetics. However, doing so induces a selection bias for those genes and gene regions whose structure is highly conserved. In turn, this means that a given column of the alignment corresponds to a site in the protein (or in the ribosomal RNA) sitting in a very specific biochemical environment inducing strong site-specific purifying selection for maintaining the conformational stability of the macromolecule (e.g. buried sites will accept only hydrophobic amino acids, exposed sites polar amino acids, etc), selection which is stable in the long run (Ashenberg et al., 2013).

In terms of the resulting sequence evolutionary process, this modulation of the selective constraint across sites will translate into a variation in both the rate and the patterns of substitutions across aligned positions. How will such a widespread variation impact phylogenetic estimation? Should we explicitly account for this variation across sites in the model used for phylogenetic reconstruction, or is it sufficient to capture the average substitution process across all sites? If explicit modeling turns out to be important for phylogenetic accuracy, then, how can we design models that will accurately capture the distribution of substitution rates and patterns across sites? These have been important questions in recent phylogenomics.

3.2 Variation of rates across sites: a parametric random-effect model

Accounting for heterogeneity in rates across sites was proposed early on, in a maximum likelihood context (see Yang (1994); Chapter 1.1 [Pupko and Mayrose 2020]). The parametric random-effect approach that was then used was subsequently ported to Bayesian inference, without major modification. It may be useful to look in detail at the conceptual structure of this approach, before addressing the more challenging question of how to model pattern-heterogeneity across sites.

The fundamental idea of the rates-across-sites model introduced by Yang (1994) is to consider site-specific relative rates as random variables, whose distribution across sites is assumed to be a gamma, of mean 1 (since these rates are relative), and of unknown variance tuned by a shape parameter noted α. Mathematically, the likelihood at site i is thus integrated over all possible values for the rate of evolution r, over a gamma distribution:

$$p(D_i \mid \theta, \alpha) \quad = \quad \int_r p(D_i \mid \theta, r) f_\alpha(r) dr,$$

where $f_\alpha(r)$ is the probability density function of the gamma distribution and, for notational simplicity, we refer to all global parameters of the model (other than α) by θ (tree topology, branch lengths, and parameters of the model of sequence evolution). In practice, this integral is intractable, and the standard approach is to numerically estimate it by discretization over a small number K of rate values $(r_k)_{k=1\ldots K}$ centered on the K quantiles of the gamma distribution (typically $K = 4$):

$$p(D_i \mid \theta, \alpha) \quad \simeq \quad \frac{1}{K} \sum_{k=1}^{K} p(D_i \mid \theta, r_k). \tag{6}$$

Finally, we can take the product over all sites:

$$p(D \mid \theta, \alpha) = \prod_i p(D_i \mid \theta, \alpha)$$

which gives the likelihood for the whole sequence alignment. This likelihood can be maximixed with respect to θ and α. Alternatively, in a Bayesian settings, one would define priors over the parameters of the model, θ and α, and then sample from the joint posterior distribution:

$$p(\theta, \alpha \mid D) \quad \propto \quad p(D \mid \theta, \alpha) p(\theta) p(\alpha).$$

Some comments and precisions are in order. First, site-specific rates are integrated over a distribution, and the distribution itself (specifically, its variance, which is equal to $1/\alpha$) is estimated across sites. In a maximum likelihood context, an alternative approach would be possible, at least in principle, namely, maximizing the likelihood with respect to all rates at all sites. However, unless the number of taxa is very large and the tree very long, this would result in overfitting. The site-specific rates would be estimated with large stochastic errors, which would then induce further estimation error on the tree topology. In addition, the variance in the rates thus estimated across sites would be greater than the variance of the true rates, since it would include the additional contribution of the error on rate estimation.

In contrast, dealing with rates as random-effects automatically discounts this additional sampling variance and returns, through α, a fair estimate of the variance of the true rates across sites. This random-effect model also achieves a higher accuracy in the estimation of the tree topology and the global parameters. This is an important point about random-effect models more generally: after fitting the model to the data, we may still have a large uncertainty about the value of the random-effects, yet, in many situations, we will nevetheless

achieve asymptotically consistent estimation of their *distribution*, and of the global parameters of the model (in particular, the tree topology). This is a recurrent idea in many other settings (e.g. integrating over gene genealogies in the multi-species coalescent [Yang 2002]).

Second, the gamma rates model considered above is a *parametric* random-effect model, since we make the assumption that the distribution of rates across sites belongs to a parametric family which is specified in advance (here, a gamma distribution). Nothing guarantees that this assumption will not be violated in practice. The true distribution of rates across sites could be arbitrary, and a mismatch between this true distribution and the distribution assumed by the model could in principle have a non-negligible impact on the accuracy of the estimation of the phylogeny. In general, it is commonly assumed that a gamma (or a mix of a proportion of invariant sites and a gamma distribution) will provide a good enough description of the true distribution of rates across sites, at least for the purpose of phylogenetic reconstruction. Of note, alternative approaches have been proposed to relax this assumption specifically for rates (Mayrose et al., 2005; Huelsenbeck and Suchard, 2007), some of which are similar to those considered below in the case of pattern-heterogeneity.

3.3 Amino acid preferences across sites: non-parametric models

As mentioned above, not just rates, but nucleotide substitution or amino acid replacement patterns, may vary across sites. In the following, and for the sake of the argument, we will more specifically consider the case of amino acid sequences. Given that the primary factor that varies across sites is selection, perhaps the most important feature whose variation across sites should be modeled is amino acid preferences, as a proxy for amino acid fitness. Mathematically, site-specific amino acid preferences can be captured through the 20-dimensional vector of amino acid equilibrium frequencies of the process. In the following, this vector will be called an amino acid frequency *profile*.

An analogy with site-specific rates suggests that we should model amino acid profiles as site-specific random effects, and also, that we should have a method allowing for a sufficiently accurate estimation of the true distribution of amino acid profiles across sites. However, there are important technical differences between site-specific rates and site-specific amino acid profiles, which are such that the method used for rates cannot be directly generalized to the present context. First, the quantile-based discretization approach mentioned above for integrating the likelihood over site-specific rates (Eq. 6) does not scale up well to higher-dimensional random-effects and would not work in practice for 20-dimensional frequency vectors. Another problem is that the true distribution of amino acid profiles is potentially complex, possibly multimodal, and thus probably not well described by any known simple parametric distribution.

Mixture models

A possible alternative is to use a finite mixture model (Koshi and Goldstein, 1998; Pagel and Meade, 2004). The rationale behind mixture models is that the diversity of the patterns of amino acid preferences realized across the aligned positions of empirical sequences might hopefully be captured by a reasonably small number of typical amino acid profiles (e.g. hydrophobic, polar, negatively charged, aromatic, etc). Allowing for K components, each with its own 20-dimensional frequency profile π_k and its own weight w_k, the likelihood at site i is then a weighted average over all mixture components:

$$p(D_i \mid \theta, \pi, w) \simeq \sum_{k=1}^{K} w_k p(D_i \mid \theta, \pi_k) \quad (7)$$

and then taking the product over all sites:

$$p(D \mid \theta, \pi, w) = \prod_{i=1}^{n} p(D_i \mid \theta, \pi, w)$$

gives the likelihood, which now depends on the set of profiles (collectively noted π) and the weight vector w, in addition to the other parameters of the model, collectively referred to as θ. In a maximum likelihood context, this likelihood will typically be maximized with respect to θ, π and w. Alternatively, the series of K profiles can be pre-estimated on a database and then kept fixed during phylogenetic inference (so-called empirical mixture models).

Much effort has been spent on deriving empirical mixtures that could be routinely used in phylogenetics (Quang et al., 2008; Le et al., 2008, 2012; Wang et al., 2014). Thus far, however, this approach has produced mixed results. One main problem is that the number of distinct profiles that seems to be required in order to obtain a good empirical fit and a sufficient phylogenetic accuracy is high (Quang et al., 2008), suggesting that the true distribution of amino acid preferences across sites might be too complex, or too diffuse, to be described by a small number of typical amino acid profiles. Practically, however, allowing for a large number of components quickly raises computational and statistical challenges, at least in a maximum likelihood framework. Computationally, averaging the likelihood at each site over all profiles of the mixture (Eq. 7) becomes prohibitive for large K. Statistically, rich mixtures quickly become redundant, in the sense that many alternative mixture configurations, differing only in small details (e.g. with several components having similar profiles), will typically give essentially equivalent approximations of the unknown empirical distribution, thus leading to poorly identifiable models.

These problems, however, are not so critical in a Bayesian framework, for two different reasons, related to the way Bayesian inference deals with model complexity (see above, Section 2). First, in a Bayesian MCMC context, parameter expansion can be used to avoid the explicit sum over all components for each site (Eq. 7). Instead, one can explicitly sample the allocations of sites to the components of the mixture during the MCMC. Combining this approach with various data-augmentation strategies allows one to design an MCMC strategy whose complexity becomes relatively insensitive to the number of components of the mixture (Lartillot, 2006). Second, the redundancy of rich mixtures, i.e. the fact that alternative mixtures effectively emulate the same distribution of random-effects, is automatically taken care of by averaging out over the posterior distribution of all possible mixture configurations.

Non-parametric random-effect models

These observations suggest that we can in fact use mixture models in a completely different regime: rather than trying to keep the number of components as low as possible, at the cost of not correctly capturing the true empirical diversity of biochemical profiles, one can instead aim for very rich and redundant mixtures. Doing so gives more flexibility. Sufficiently rich mixtures can approximate any distribution with arbitrary accuracy, and the fact that they are redundant does not matter so much, as long as an efficient MCMC is able to smooth out this redundancy by averaging over a representative sample of alternative mixture configurations, all of which giving essentially equivalent approximations of the true distribution. This is the fundamental idea behind Bayesian non-parametric random-effect models.

The original goal of non-parametric inference is to relax the assumption that the true distribution should a priori belong to a pre-specified parametric family. In principle, a non-parametric approach should give asymptotically consistent results for arbitrary distributions of random effects across sites. In a Bayesian context, this is implemented by designing a prior

over rich mixtures. The Dirichlet process is such a non-parametric prior (Ferguson, 1973; Müller and Mitra, 2013). Technically, the Dirichlet process pushes the idea of sufficiently rich mixtures to its extreme, by implementing a prior over *infinite* mixtures. Infinite mixtures are dense in the space of all possible distributions, and thus, a Dirichlet process prior will put some probability mass in the vicinity of any distribution – including of course the true distribution of random-effects across sites. Then, conditioning the model on a sufficiently large dataset will result in a posterior distribution which will concentrate in the vicinity of the true distribution. In the end, implementing this idea using clever MCMC approaches based on parameter expansion will effectively implement a powerful non-parametric inference method, in principle achieving asymptotic consistency under arbitrary distributions of (possibly multi-dimensional) random-effects.

Dirichlet process priors have been applied to several problems in phylogenetics, for modeling variation across sites in rates (Huelsenbeck and Suchard, 2007), dN/dS (Huelsenbeck et al., 2006), amino acid preferences across sites (Lartillot and Philippe, 2004) or amino acid fitness profiles in the context of mechanistic mutation-selection codon models (Rodrigue et al., 2010); but also, for modeling variation in rates across branches in a relaxed clock model (Heath et al., 2012).

4 Software programs and platforms

A large number of software programs are currently available for conducting phylogenetic or phylogeny-related Bayesian inference. These programs often have very different specific objectives or specializations: phylogenetic reconstruction (Ronquist and Huelsenbeck (2003); Lartillot et al. (2013); Lewis et al. (2015); Chapter 1.5 [Lartillot 2020]), molecular dating (Thorne and Kishino (2002); Rannala and Yang (2007); Chapter 5.2 [Barido-Sottani et al. 2020]), phylogeography and phylodynamics (Bouckaert et al., 2014), phylogenetic codon models (Murrell et al. (2013); Rodrigue and Lartillot (2014); Chapter 4.5 [Lowe and Rodrigue 2020]), comparative studies (Pagel et al., 2004), gene-tree species-tree reconciliation (Akerborg et al., 2009), or species delimitation (Yang and Rannala (2010); Chapter 5.6 [Flouri et al. 2020]).

In spite of this current move toward integrative modeling approaches (and perhaps in part because of the computational challenges), much of current applied research in phylogenomics is still conducted in the context of the more classical supermatrix paradigm, in which a large set of single-gene alignments are simultaneously considered and assumed to evolve along the same species tree. Three main software programs have been used in recent phylogenomic analyses more specifically devoted to reconstructing a species tree using supermatrices:

- MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) is the most widely used program for Bayesian phylogenetic reconstruction. It offers a broad range of models, allowing for standard nucleotide, amino acid and codon models, but also models of morphological character evolution, while offering the possibility to analyze heterogeneous datasets (i.e. mixing morphological, DNA or RNA and amino acid data). On the other hand, this program has limited expressivity for pattern-heterogeneity across sites within partitions. MrBayes has been extensively used for standard phylogenetic analysis, and in several recent phylogenomic studies (Cannon et al., 2016).
- PhyloBayes (Lartillot et al., 2009, 2013) is a program specialized in site-heterogeneous models of sequence evolution. Its main distinguishing feature is the use of Dirichlet process priors, such as introduced above, to model the variation across sites in amino acid preferences. In part because of the increasing awareness of the importance of accounting

for site-heterogeneity for reconstructing ancient phylogenies, this program has been increasingly used over the recent years, in particular for reconstructing the metazoan tree of life (Simion et al., 2017), but also eukaryotes (Brown et al., 2018), Archaea (Adam et al., 2017), or Eubacteria (Antunes et al., 2016). A detailed application using PhyloBayes is presented in Chapter 1.5 of this book (Lartillot, 2020).

- ExaBayes (Aberer et al., 2014) is a recent implementation of Bayesian phylogenetic inference for very large supermatrices. This program allows for heterogeneity across partitions, as well as rate-heterogeneity (but not pattern-heterogeneity) across sites within partitions.
- P4 (Foster, 2004) is specialized in branch-heterogeneous models of sequence evolution, more specifically, accounting for compositional heterogeneity across taxa. This program has been used, in particular, for investigating the position of eukaryotes in the tree of life (Cox et al., 2008).

Ideally, it would be very useful to have a single implementation combining these various levels of expressiveness in model design (i.e. allowing for gene-, site- and branch-specific modulations in both rate and pattern heterogeneity), all of which appear to be essential in order to achieve accurate phylogenetic reconstruction. Thus far, however, no such integrated implementation is available. One main reason for this vacancy is the computational complexity inherent to each of these multiple sources of variation, which would be compounded in a joint implementation and further aggravated by the size of the current datasets in phylogenomics.

5 Conclusions and perspectives

In several respects, Bayesian inference has revolutionized our practice in phylogenetics, although perhaps not for the reasons that have often been invoked. In theory, Bayesian inference offers a flexible framework for introducing subjective or context information through the prior. In practice, however, this is not the main reason behind the recent success and popularity of Bayesian inference in evolutionary genetics. Instead, it is the combination of hierarchical models and generic Monte Carlo approaches for dealing with complex random-effects and multi-level evolutionary processes that has played the most important role.

Modeling pattern-heterogeneity across sites represents one specific instance where complex random-effect models turn out to have an important impact on practical phylogenomics. This problem is challenging in two respects: first, because the random effects to be modeled (amino acid preferences) are high-dimensional, and second, because the distribution of those random-effects across sites is itself unknown and apparently complex. This combination makes Bayesian inference using Monte Carlo particularly well-suited, whereas simpler approaches, such as parametric random-effect models or finite mixture models, have thus far shown less affordable or less accurate.

However, at least given the current state-of-the art, Bayesian inference suffer from several limitations. First, current Monte Carlo algorithms do not scale up well with data size, and as a result, Bayesian inference can become computationally prohibitive for large phylogenomic datasets. This is particularly true for non-parametric models based on Dirichlet process priors. Second, the flexibility afforded by Bayesian inference for handcrafting new and complex multi-level models is nice in theory, yet in practice, it requires a substantial amount of programming work for each new model that one might want to contemplate. This also raises the question of the reliability of the software implementations, as it is typically difficult to guarantee that a given implementation of a Monte Carlo algorithm is indeed sampling from the intended target distribution.

In this direction, the current trend is in the development of generic programming platforms. The fact that model components can be composed like building blocks into complex hierarchical structures, using modular Monte Carlo methods to explore the resulting posterior distribution, makes generic programming for Bayesian inference relatively straightforward, at least conceptually. Such generic programming platforms have been proposed both in general applied statistics (Lunn et al., 2009) and more recently in phylogenetics (Höhna et al. 2016; Chapter 5.4 [Ayres et al. 2020]). They represent a promising development, by providing the applied evolutionary scientific community with user-oriented tools for reliably designing question-specific integrative models and applying them to arbitrary combinations of empirical data (genetic sequences, morphological data, time series, etc). The computational challenges, however, are formidable, and much remains to be done in Monte Carlo algorithmics and software development, in order for Bayesian generic programming to achieve scalable inference in the context of current problems in evolutionary genomics.

References

Aberer, A. J., Kobert, K., and Stamatakis, A. (2014). ExaBayes: massively parallel bayesian tree inference for the whole-genome era. *Mol. Biol. Evol.*, 31(10):2553–2556.

Adam, P. S., Borrel, G., Brochier-Armanet, C., and Gribaldo, S. (2017). The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J*, 11(11):2407–2425.

Akerborg, O., Sennblad, B., Arvestad, L., and Lagergren, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. USA*, 106(14):5714–5719.

Antunes, L. C., Poppleton, D., Klingl, A., Criscuolo, A., Dupuy, B., Brochier-Armanet, C., Beloin, C., and Gribaldo, S. (2016). Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *Elife*, 5.

Ashenberg, O., Gong, L. I., and Bloom, J. D. (2013). Mutational effects on stability are largely conserved during protein evolution. *Proc. Natl. Acad. Sci. USA*, 110(52):21071–21076.

Ayres, D. L., Lemey, P., Baele, G., and Suchard, M. A. (2020). Beagle 3 high-performance computational library for phylogenetic inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.4, pages 5.4:1–5.4:9. No commercial publisher | Authors open access book.

Barido-Sottani, J., Justison, J. A., Wright, A. M., Warnock, R. C. M., and Pett, W. (2020). Estimating a time-calibrated phylogeny of fossil and extant taxa using revbayes. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.2, pages 5.2:1–5.2:23. No commercial publisher | Authors open access book.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.*, 10(4):e1003537.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Bromham, L. (2020). Substitution rate analysis and molecular evolution. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.4, pages 4.4:1–4.4:21. No commercial publisher | Authors open access book.

Brown, M. W., Heiss, A. A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A. K., Shiratori, T., Ishida, K.-I., Hashimoto, T., Simpson, A. G. B., and Roger, A. J. (2018). Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biol. Evol.*, 10(2):427–433.

Cannon, J. T., Vellutini, B. C., Smith, J., Ronquist, F., Jondelius, U., and Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, 530(7588):89–93.

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., and Embley, T. M. (2008). The archaebacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. USA*, 105(51):20356–20361.

De Finetti, B. (1974). *Theory of Probability.* John Wiley and Sons, New York.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230.

Flouri, T., Rannala, B., and Yang, Z. (2020). A tutorial on the use of bpp for species tree estimation and species delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.6, pages 5.6:1–5.6:16. No commercial publisher | Authors open access book.

Foster, P. G. (2004). Modeling Compositional Heterogeneity. *Syst. Biol.*, 53(3):485–495.

Heath, T. A., Holder, M. T., and Huelsenbeck, J. P. (2012). A dirichlet process prior for estimating lineage-specific substitution rates. *Mol. Biol. Evol.*, 29(3):939–955.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc. Natl. Acad. Sci. USA*, 111(29):E2957–66.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27(3):570–580.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Syst. Biol.*, 65(4):726–736.

Holder, M. and Lewis, P. (2003). Phylogenetic estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.*, 4:275–284.

Huelsenbeck, J., Rannala, B., and Masly, J. (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, 288(5475):2349–2350.

Huelsenbeck, J. P., Jain, S., Frost, S. W. D., and Pond, S. L. K. (2006). A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl. Acad. Sci. USA*, 103(16):6263–6268.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Huelsenbeck, J. P. and Suchard, M. A. (2007). A nonparametric method for accommodating and testing across-site rate variation. *Syst. Biol.*, 56(6):975–987.

Koshi, J. M. and Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins*, 32(3):289–295.

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. (2014). Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *Journal of The Royal Society Interface*, 11(94):20131106.

Larget, B. and Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.*, 16:750–759.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *J. Comput. Biol.*, 13(10):1701–1722.

Lartillot, N. (2020). Phylobayes: Bayesian phylogenetics using site-heterogeneous models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.5, pages 1.5:1–1.5:16. No commercial publisher | Authors open access book.

Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–2288.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, 21(6):1095–1109.

Lartillot, N. and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.*, 28(1):729–744.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, 62(4):611–615.

Le, S. Q., Dang, C. C., and Gascuel, O. (2012). Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.*, 29(10):2921–2936.

Le, S. Q., Lartillot, N., and Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 363(1512):3965–3976.

Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. (2007). A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.*, 24(12):2669–2680.

Lewis, P. O., Holder, M. T., and Swofford, D. L. (2015). Phycas: software for Bayesian phylogenetic analysis. *Syst. Biol.*, 64(3):525–531.

Li, S., Pearl, D. K., and Doss, H. (2000). Phylogenetic Tree Construction Using Markov Chain Monte Carlo. *J. Am. Stat. Assoc.*, 95(450):493–508.

Lowe, C. and Rodrigue, N. (2020). Detecting adaptation from multi-species protein-coding dna sequence alignments alignments. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.5, pages 4.5:1–4.5:18. No commercial publisher | Authors open access book.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.

Mau, B., Newton, M. A., and Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55(1):1–12.

Mayrose, I., Friedman, N., and Pupko, T. (2005). A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, 21 Suppl 2:ii151–8.

Müller, P. and Mitra, R. (2013). Bayesian Nonparametric Inference - Why and How. *Bayesian Analysis*, 8(2).

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., and Scheffler, K. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.*, 30(5):1196–1205.

Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, 53(4):571–581.

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.*, 53(5):673–684.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Quang, L. S., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Rannala, B. and Yang, Z. (2007). Inferring speciation times under an episodic molecular clock. *Syst. Biol.*, 56(3):453–466.

Rodrigue, N. and Lartillot, N. (2014). Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics*, 30(7):1020–1021.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. USA*, 107(10):4629–4634.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, E., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., and Manuel, M. (2017). A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol.*, 27(7):958–967.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Thorne, J. L. and Kishino, H. (2002). Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.*, 51(5):689–702.

Wang, H.-C., Susko, E., and Roger, A. J. (2014). An amino acid substitution-selection model adjusts residue fitness to improve phylogenetic estimation. *Mol. Biol. Evol.*, 31(4):779–792.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39(3):306–314.

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4):1811–1823.

Yang, Z. (2007). Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. *Mol. Biol. Evol.*, 24(8):1639–1655.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol. Biol. Evol.*, 14(7):717–724.

Yang, Z. and Rannala, B. (2006). Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.*, 23(1):212–226.

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. USA*, 107(20):9264–9269.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. (2016). Total-Evidence Dating under the Fossilized Birth-Death Process. *Syst. Biol.*, 65(2):228–249.

Zhukova, A., Gascuel, O., Duchêne, S., Ayres, D. L., Lemey, P., and Baele, G. (2020). Efficiently analysing large viral data sets in computational phylogenomics. In Scornavacca,

C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.3, pages 5.3:1–5.3:43. No commercial publisher | Authors open access book.

Chapter 1.5 PhyloBayes: Bayesian Phylogenetics Using Site-heterogeneous Models

Nicolas Lartillot
Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive,
43 Bld du 11 Novembre 1918, 69622 Villeurbanne cedex, France.
nicolas.lartillot@univ-lyon1.fr
https://orcid.org/0000-0002-9973-7760

Abstract

PhyloBayes is a software program for Bayesian phylogenetic reconstruction. Compared to other programs, its main distinguishing feature is the implementation of the CAT model, which accounts for fine-grained variation across sites in amino acid preferences using a Bayesian non-parametric approach. This chapter provides a detailed step-by-step practical introduction to phylogenetic analyses using PhyloBayes, using as an example a previously published dataset addressing the phylogenetic position of Microsporidia within eukaryotes. Through this historically emblematic case of a long-branch attraction artifact, a complete analysis under site-homogeneous and site-heterogeneous models is conducted and interpreted, thus providing an illustration of why modeling pattern variation is so fundamental for reconstructing deep phylogenies.

How to cite: Nicolas Lartillot (2020). PhyloBayes: Bayesian Phylogenetics Using Site-heterogeneous Models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 1.5, pp. 1.5:1–1.5:16. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Introduction

Since the realization, by Zuckerkandl and Pauling (1965), that DNA molecules represent documents of the long-term evolutionary history of species, molecular phylogenetics has gone a long way in progressively deciphering the detailed patterns of diversification across species at all evolutionary scales. Accurately reconstructing the tree of life, however, has turned out to be quite more challenging than anticipated, especially over deep evolutionary times. The long-standing hesitations, over the last 50 years, concerning the position of Microsporidia in the tree of eukaryotes (Brinkmann et al., 2005), or that of nematodes (Aguinaldo et al., 1997; Philippe et al., 2005) and, more recently, ctenophores (Telford et al., 2016), in the metazoan phylogeny, clearly illustrate the difficulty in firmly establishing a definitive picture of the diversification patterns having occurred in the remote evolutionary past.

The first phylogenetic methods, based on distance or on maximum parsimony, have quickly shown important methodological weaknesses and have progressively been replaced by more principled model-based approaches, using either maximum likelihood or Bayesian inference. Even with these probabilistic methods, however, are systematic errors in tree reconstruction still a major plague (Philippe et al. 2011; Chapter 2.1 [Simion et al. 2020]). One main reason is the difficult question of model adequacy: probabilistic approaches are accurate only inasmuch as the underlying model of sequence evolution correctly describes the true evolutionary process. In practice, models are obviously a much idealized description of the true processes, which thus raises the question of which aspects of the evolutionary process are critical and should be correctly captured, in order to mitigate the impact of reconstruction errors.

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 1.5; pp. 1.5:1–1.5:16

A book completely handled by researchers.
No publisher has been paid.

In this direction, variation among sites turns out to be a particularly important feature to take into account. The simple models originally used in phylogenetics typically assume that all sites evolve under the same process of nucleotide or amino acid substitutions. In the case of amino acid sequence alignments, such models are typically implemented using so-called empirical amino acid replacement matrices, such WAG (Whelan and Goldman, 2001) or LG (Le and Gascuel, 2008). This, however, amounts to assuming that all sites should visit amino acid states at the same relative frequencies. Yet in practice, there is much variation among sites in substitution patterns (and in particular, in amino acid preferences). As it turns out (and as will be explored in more detail below), explicitly accounting for this heterogeneity across sites is crucial, in order to get more accurate tree reconstructions.

Owing to its complexity, pattern variation across sites is not a trivial aspect of the evolutionary process to model adequately. This problem has motivated (and is still motivating) the development of various approaches (Halpern and Bruno, 1998; Koshi and Goldstein, 1998; Lartillot and Philippe, 2004; Pagel and Meade, 2004; Wang et al., 2008; Le et al., 2008; Quang et al., 2008; Wang et al., 2014, 2018; Susko et al., 2018; Dang and Kishino, 2019). In particular, the CAT model (Lartillot and Philippe, 2004), such as implemented in PhyloBayes (Lartillot et al., 2013), relies on a Bayesian non-parametric approach based on Dirichlet process priors (see Chapter 1.4 [Lartillot 2020] for an introduction on the concepts of Bayesian inference, site-heterogeneity and non-parametric models). In this chapter, a practical application using PhyloBayes is presented, showing how to use both site-homogeneous and site-heterogeneous model, compare their results and evaluate their goodness of fit. The results are then interpreted in the broader context of phylogenomic analysis over broad evolutionary scales.

2 A practical example using PhyloBayes: Microsporidia

As a practical example of how to conduct a Bayesian phylogenetic analysis with PhyloBayes, we consider here a phylogenomic dataset originally assembled by Brinkmann et al. (2005). This dataset is a concatenation of 133 genes (24,000 aligned positions) for 40 taxa (34 eukaryotes and 6 Archaea). It represents an interesting case, for which the inferred position of the fast-evolving Microsporidia in the phylogeny of eukaryotes turns out to be model-dependent – and, more specifically, turns out to depend on whether or not site-specific amino acid preferences are accounted for.

All analyses presented here have been conducted using PhyloBayes MPI, version 1.8. The package can be obtained directly from github (`https://github.com/bayesiancook/pbmpi`). The dataset is also available along with the current version of the program. In what follows, we briefly recall the main points about program usage that are necessary to run through the complete analysis on this particular dataset. For more information, see the manual (provided in the package).

2.1 Running PhyloBayes under the CAT model

PhyloBayes is primarily intended for high-performance computing facilities operating under linux or unix. The package contains a series of programs (`pb_mpi`, `readpb_mpi`, `bpcomp`, `tracecomp`), all of which can be controlled using a command-line interface. Among them, `pb_mpi` implements the Markov chain Monte Carlo sampler targeting the posterior distribution over the parameters of the model chosen by the user. The MCMC sampler cycles over a complex series of Monte Carlo updates (or moves) of the topology, the branch lengths or the substitution model (including the Dirichlet process mixture), and saves the current model

configuration after each cycle. The series of points saved during a run of `pb_mpi` defines a *chain*. Each chain has a name, which is used as the base name for all files produced during the run.

Running a chain using `pb_mpi`

To start our analysis, we run a first chain under the CAT-F81 model on the Microsporidia dataset. This model, which combines uniform exchange rates across amino acid pairs (Felsenstein, 1981, generalized to amino acid states) with site-specific amino acid equilibrium frequency profiles from a Dirichlet process (the CAT model), was introduced by Lartillot and Philippe (2004) and represents the best compromise between computational speed and phylogenetic robustness. Runs under this model are much faster than under alternative models, such as considered below. Therefore, it is generally useful to start with CAT-F81, so as to get a first picture of the problem of interest, before launching computationally more demanding analyses. To run the chain, we type the following command:

```
mpirun -np 32 pb_mpi -d microsporidia.ali -cat -f81 -dgam 4 catf81microspo1 &
```

Here, we have started the analysis in direct mode. On a cluster operated by a job scheduling system, one would instead need to write a script containing, among other information, the command for running `pb_mpi`, and then send it to the queue.

In this command, the `-np 32` option specifies the number of processes running in parallel (this number should be at least 2). The `-d` option is for specifying the dataset. For the model, we combine a Dirichlet process for site-specific equilbirium frequency profiles over amino acids (the `-cat` option) with uniform (or Poisson) exchangeabilities (the `-f81` option). In addition, we allow for rate variation across sites, using a discretized gamma distribution with 4 categories (`-dgam 4`). Finally, we give a name to the chain, here, `catf81microspo1`. Before starting, the chain will output a summary of the model settings.

While the chain is running, a series of files will be produced. The most important are:

- `catf81microspo1.treelist`: list of sampled trees (with branch lengths);
- `catf81microspo1.trace`: the trace file, containing summary statistics (detailed below);
- `catf81microspo1.chain`: contains the parameter configurations visited during the run.

These files will be regularly updated (after each cycle). Note that the trace file contains one line per point saved since the beginning of the run. Thus, the number of lines of the trace file gives a direct indication of the current MCMC sample size. It is always good practice to run at least two chains in parallel and compare the samples obtained under these several independent runs. In the present case, we run four independent chains, which we name `catf81microspo1`, `2`, `3`, and `4`.

The chains will run as long as allowed. PhyloBayes implements a check-pointing system, so that chains can be interrupted at any time (possibly because of a timeout on the cluster) and then restarted. In the present case, on the machine where the example was conducted, the four independent chains save one point about every 30 seconds, or 150 points per hour. We let these chains run for 24 hours ($\sim$ 3500 points) before checking the results.

Checking convergence and mixing (`bpcomp` and `tracecomp`)

Convergence can be first visually assessed by plotting the summary statistics recorded in the trace files as a function of number of iterations. Visual assessment can be conducted while the chain is running. This can be done on the fly, directly from the command line interface, using simple linux utilities such as `gnuplot`. Alternatively, the trace file of `PhyloBayes` is

1.5:4 Bayesian phylogenetics

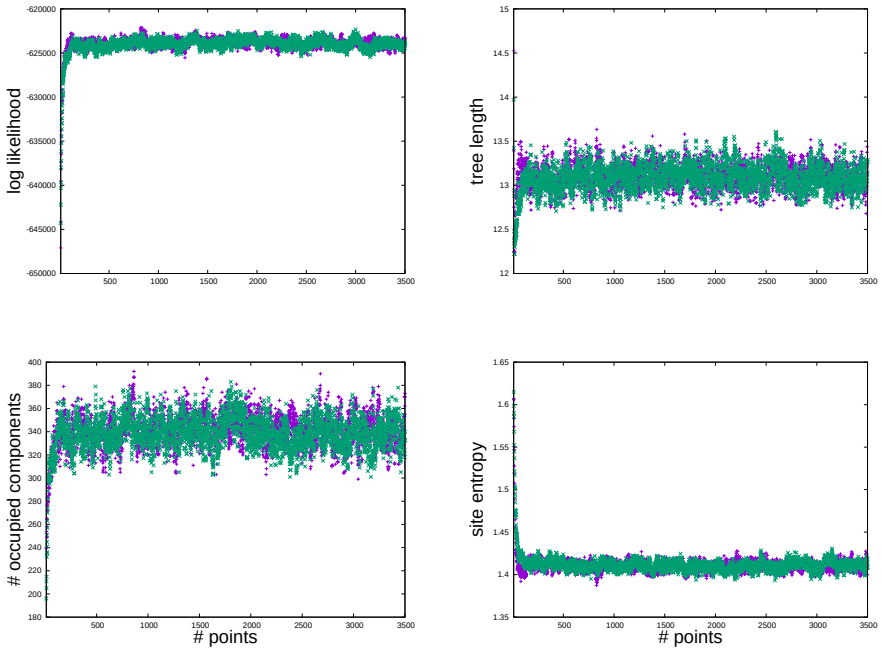


Figure 1 Traceplots for the CAT analyses, showing, as a function the number of points saved in the tracefile, the log likelihood, the tree length, the number of occupied components and the mean site entropy

compatible with the `Tracer` program (Rambaut et al., 2018). Visual assessment is essential, in particular, for gettting a reliable estimate of the burn-in, i.e the number of points before the chain has reached stationarity. In general, it is particularly important to visualize at least the log likelihood (`loglik`, 4th column of the trace file), the total tree length (`length`, column 5), the number of occupied components of the mixture (`Nmode`, column 6) and the mean site entropy (`statent`, column 7), which is a measure of the strength of site-specific amino acid preferences. In the present case, after 24 hours, the four independent chains have saved around 3500 points each. Visualization of the log likelihood and the summary statistics (Figure 1) suggests that the chains have reached convergence after a burnin of 400 to 500 points. We set the burnin conservatively to 500.

After visual inspection, convergence and mixing can be assessed more quantitatively. This can be done using the `tracecomp` program (for checking convergence of the parameters) and the `bpcomp` program (for assessing convergence in tree space). Both use a similar syntax. First, we inspect the trace files using tracecomp:

```
tracecomp -x 500 catf81micro1 catf81micro2 catf81micro3 catf81micro4
```

or, more rapidly

```
tracecomp -x 500 catf81micro?.trace
```

which produces an output summarizing the estimated effective sample size and the discrepancies among the four runs for each column of the trace file:

```
name                effsize     rel_diff
```

```
loglik          101     0.302739
length          614     0.0754705
alpha           611     0.0914714
Nmode           245     0.177897
statent         257     0.126228
statalpha       631     0.317062
kappa           471     0.120433
```

The effective size (second colum, `effsize`) is an estimate of the effective number of independent points produced by each run. As for the discrepancy (third column, `rel_diff`) it measures, for each statistic of the trace file, the deviation among the four chains in the mean value, normalized by the within-chain standard deviation of the statistic and averaged over all pairs of runs. A discrepancy much less than 1 means that the error on the posterior mean estimate for a given quantity is very small compared to the 95% credible interval. Here, all effective sizes are greater than 100 (that is, each chain yields the equivalent of at least 100 independent draws from the posterior distribution), and the discrepancies are less than or slightly above 0.3. The run is thus quite acceptable. Ideally, one would like to achieve effective sample sizes more in the order of 1000 or more, and discrepancies smaller than 0.1. However, this has typically been difficult to achieve in Bayesian phylogenomics, for reasonably large datasets. In the present case, the chains could be run for an additional few days, and the discrepancies would then decrease, eventually landing below 0.1 for all entries of the trace file. Of note, the differences that are implied by these discrepancies are relatively small in practice. For instance, in the case of `alpha` (the α parameter of the gamma distribution of rates across sites), which is the entry with the highest discrepancy (0.32), the 4 posterior mean estimates obtained for the 4 chains are all between 0.64 and 0.67 – thus implying very similar distributions of rates across sites. The reason why the discrepancies look large is that the standard deviation within each chain is small (around 0.03). Thus, the criterion of having all discrepancies below 0.1 is in fact fairly stringent.

Second, we inspect tree lists using `bpcomp`:

```
bpcomp -x 500 catf81micro1 catf81micro2 catf81micro3 catf81micro4
```

or, more rapidly:

```
bpcomp -x 500 catf81micro?.treelist
```

The program writes in the log the largest (`maxdiff`) and mean (`meandiff`) discrepancy observed across all bipartitions:

```
maxdiff     : 0.186092
meandiff    : 0.00323931

bipartition list in : bpcomp.bplist
consensus in        : bpcomp.con.tre
```

In the present case, the maximum difference in bipartition frequencies among the four chains is above 0.1 (equal to 0.186). Detailed inspection of the `bpcomp.bplist` file, however, shows that this is due only to discrepancies between chains about the position of *Glomus* within Fungi. For all other bipartitions, the discrepancy between the chains is below 0.1. Of note, a `maxdiff` of 1, which happens not so rarely in practice, warns us that at least one clade is inferred with a posterior probability of 1 in one chain and 0 in another chain, a clear sign of

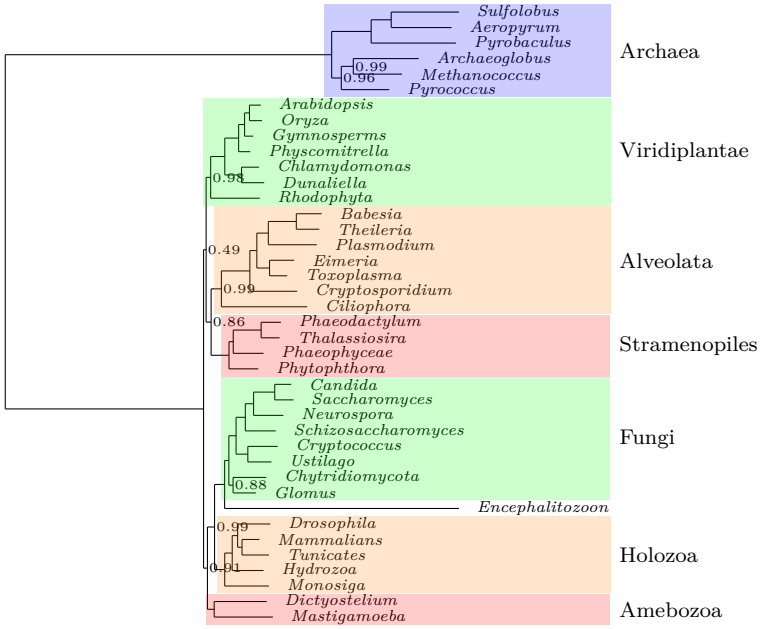


Figure 2 Consensus tree under the CAT-F81 model

MCMC mixing problems. In practice, however, as long as the discrepancies between chains does not directly affect the group of interest, this has usually been considered as acceptable.

Posterior consensus tree

The `bpcomp` program also produces a file containing the consensus obtained by pooling the trees of all of the chains given as arguments (file named `bpcomp.con.tre`). We see from this tree (Figure 2) that CAT-F81 infers that Microsporidia are the sister-group to Fungi, which is in accordance with the currently accepted view (Brinkmann et al., 2005). There is some uncertainty in the early splits at the base of eukaryotes, with posterior probability support values often smaller than 0.95. In fact, most of this lack of support is due to some hesitation about the branching point for the outgroup (Archaea). In the consensus displayed here, it is between Unikonta (Holozoa, Fungi, Amebozoa) and Bikonta (Viridiplantae, Alveolata and Stramenopiles), although this specific rooting point has a posterior probability of only 0.49.

2.2 Running PhyloBayes under other models

We can now run a similar analysis under other models. Here, we consider CAT-GTR, LG (Le and Gascuel, 2008) and GTR. The commands are as follows (in each case, with 4 replicates):

```
mpirun -np 32 pb_mpi -d microsporidia.ali -cat -gtr -dgam 4 catgtrmicrospo1 &
mpirun -np 32 pb_mpi -d microsporidia.ali -ncat 1 -gtr -dgam 4 gtrmicrospo1 &
mpirun -np 32 pb_mpi -d microsporidia.ali -ncat 1 -lg -dgam 4 lgmicrospo1 &
```

Note that the command-line option to select one-matrix models in PhyloBayes is just `-ncat 1`, that is, a mixture with one single category (one single matrix for all sites).

For these three models, the time per cycle is substantially longer than under CAT-F81: 1 point per minute for GTR and LG, and 1 point every 2 minutes for CAT-GTR. For the

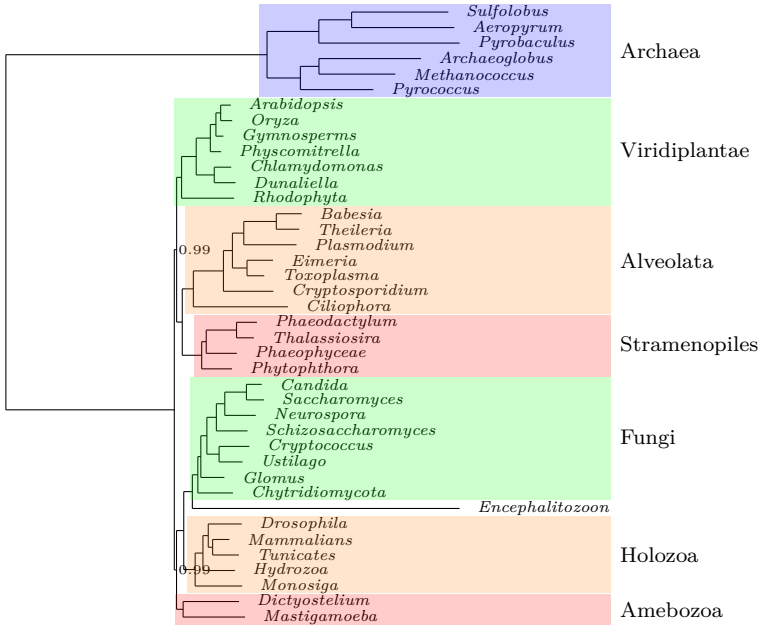


Figure 3 Consensus tree under the CAT-GTR model

CAT-GTR model, we will thus need 5 to 6 days in order for our chains to reach a size of 3500. Checking convergence on CAT-GTR after 5 days gives results similar to those obtained for CAT-F81. With `tracecomp`, effective sizes are all greater than 100. Discrepancies, on the other hand, are a bit larger between independent chains than what was observed with CAT-F81, although still acceptable:

```
name                effsize     rel_diff

loglik              374         0.434818
length              101         0.159781
alpha               980         0.178217
Nmode               466         0.123387
statent             241         0.38135
statalpha           291         0.126557
kappa               704         0.0996022
rrent               1005        0.21717
rrmean              2852        0.0308481
```

With `bpcomp`, reproducibility between chains is high:

```
maxdiff      : 0.0219639
meandiff     : 0.000298828
```

The resulting consensus tree (Figure 3) differs from that obtained under CAT-F81 only for the position of *Glomus*, although this might be due to the lack of convergence of the CAT-F81 analyses concerning the position of this particular taxon. Another remarkable difference, compared to CAT-F81, is the higher posterior probability support values obtained by CAT-GTR for the deep clades of the eukaryotic ingroup. The most probable rooting for eukaryotes

is, again, between Unikonta and Bikonta, although now with a posterior probability greater than 0.95. This pattern is often observed when comparing these 2 models: typically, CAT-F81 tends to be more conservative than CAT-GTR, giving lower clade posterior probabilities. Otherwise, the two models do not differ so much in their point estimates.

With LG and GTR, mixing turns out to be challenging, with different chains stabilizing at different levels. This is clearly detected both by `tracecomp` and `bpcomp`. First, the `tracecomp` output points to a very high discrepancy for the log likelihoods between chains, with a `rel_diff` statistic above 25:

```
name                effsize     rel_diff

loglik              1783        25.8753
length              658         2.24296
alpha               966         1.05931
Nmode               2400        0
statent             943         0.253293
statalpha           2400        0
rrent               761         0.263397
rrmean              1991        0.0536239
```

As for `bpcomp`, it gives us a maximal discrepancy of 1:

```
maxdiff     : 1
meandiff    : 0.025974
```

A closer look at the trace files shows that the log likelihood stabilizes for all chains, after less than 100 cycles. However, some chains stabilize at -727300, while other chains reach only -727600. Obviously, the latter are stuck in a local optimum in tree space. The difference in log likelihood between chains is so large in the present case that we can fairly exclude the chains stabilizing at -727600, as not representative of the true equilibrium, and concentrate on those chains that reach the highest average log likelihood. Restricting the `tracecomp` analysis to those chains gives much better convergence statistics:

```
name                effsize     rel_diff

loglik              1586        0.00714185
length              942         0.10133
alpha               505         0.180445
Nmode               2400        0
statent             389         0.139412
statalpha           2400        0
rrent               311         0.0898343
rrmean              1877        0.0254995
```

Running `bpcomp` on those chains gives a `maxdiff` of 0, indicating that they have stabilized on one single tree. Importantly, this tree differs from the ones obtained under CAT-F81 or CAT-GTR, in that Microsporidia now appear as being the sister-group to all other eukaryotes (Figure 4). In fact, this tree differs more specifically from the tree obtained by CAT-GTR by rotating the eukaryotic ingroup, so as to root it on Microsporidia. Equivalently, if we ignore the rooting, we might see this tree as the result of Archaea moving up in the tree and becoming the sister-group to Microsporidia. The same tree topology, with maximal support,

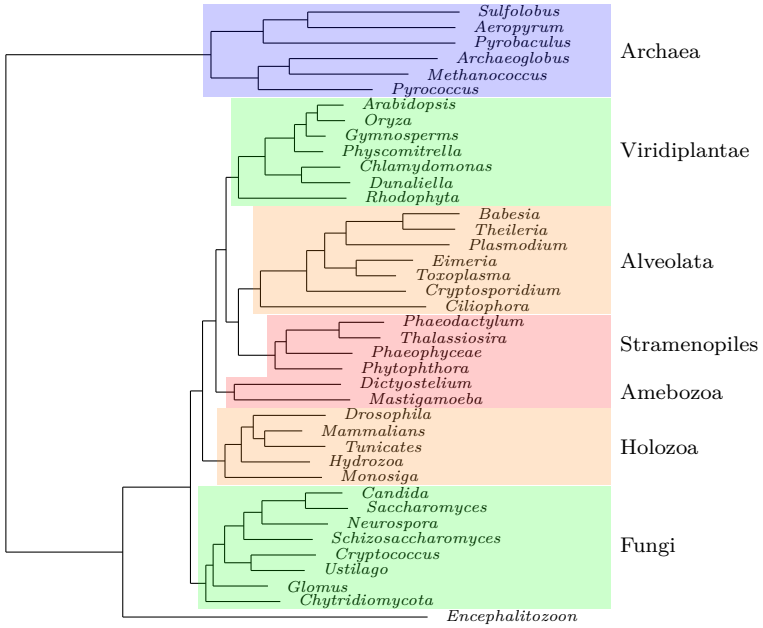


Figure 4 Consensus tree under the GTR model

is also obtained under LG (not shown). Of note, this result is confirmed by maximum likelihood analyses –such as conducted using RaxML (Stamatakis et al., 2005) for the LG model, or IQTree (Nguyen et al., 2015) for LG and GTR– which all give the topology obtained here using PhyloBayes under these 2 site-homogeneous models (not shown). This illustrates the fact that what determines the outcome of the analysis is not, in itself, the choice between maximum likelihood or Bayesian inference, but instead, the choice of the model of sequence evolution.

2.3 A typical case of long-branch attraction artifact

The analysis conducted above represents a situation where different models give different tree topologies based on the same dataset. On one side, site-homogeneous models (GTR, LG) give Microsporidia sister-group to all other eukaryotes (Figure 4); on the other side, site-heterogeneous models (CAT-F81, CAT-GTR) give Microsporidia sister-group to Fungi (Figures 2 and 3). Thus, at least one category of models is subject to a systematic error.

In the present case, we happen to have a relatively good independent knowledge of which of the two estimates obtained above is more likely to be correct. The position of Microsporidia in the eukaryotic tree is a well-known phylogenetic problem, having received a lot of attention since the early days of molecular phylogenetics. Based on the first universal trees reconstructed using 16S rRNA (Woese et al., 1990), it was at originally believed that Microsporidia were “early-emerging” eukaryotes, as also suggested here by GTR or by LG. However, there are now quite a few lines of evidence suggesting that Microsporidia are in fact closely related to Fungi (Brinkmann et al., 2005), as suggested by CAT-F81 and CAT-GTR.

In most practical situations, however, independent knowledge about the true tree is lacking. A case in point is the position of Ctenophores in the metazoan tree, which has received a lot of attention recently, and for which alternative hypotheses are still the subject

of some controversy (Telford et al., 2016). In such situations, independent and objective arguments are needed, in order to determine which of the alternative models is more likely to return an accurate estimate of the phylogeny.

In the case of Microsporidia, and assuming no independent knowledge, it can be noted that the two longest branches in the tree are, first, the branch leading to Microsporidia, and second, the branch between Eukaryotes and Archaea. The tree inferred by GTR is thus exactly what one would expect as the result of an artifactual attraction between these two long branches. Such arguments in terms of long-branch attraction artifacts are often useful. In many situations, they offer a good heuristic for making sense of the observed incompatibilities between the trees returned by alternative models, or the instability in tree estimation caused by varying the taxonomic sampling. These heuristic arguments, however, should be complemented by more formal model evaluation. This can be done along two main directions: model comparison and model checking.

2.4 Model comparison

Model comparison consists of measuring how well each model fits the data at hand. Choosing the best-fitting model (and the corresponding phylogenetic estimate) usually defines a good decision procedure, in particular if the difference in fit between alternative models is very large. In Bayesian inference, a classical measure of the fit of a model is its marginal likelihood, which is the likelihood averaged over the prior. The higher the marginal likelihood, the more likely it is that the model would produced the observed data, upon drawing a parameter configuration from its prior and then simulating the sequence evolutionary process. When comparing two models, it is customary to define the Bayes factor between these two models as the ratio of their marginal likelihoods (Jeffreys, 1935; Kass and Raftery, 1995; Gelman et al., 2004).

Marginal likelihoods have several drawbacks, however. First, on a theoretical ground, they are sensitive to the prior, and much more so than the posterior distribution itself. Second, on a more practical note, marginal likelihoods are notoriously difficult to numerically evaluate.

For those reasons, an alternative approach, used in PhyloBayes, is cross-validation. The idea of cross-validation is relatively simple: the data D are split into two subsets of unequal size, D_1 (typically, 90% of the original dataset) and D_2 (10% of the original dataset). The model is trained on D_1, and then the trained model is used to "predict" dataset D_2. In a Bayesian framework, this translates into averaging the likelihood of D_2 under the posterior distribution obtained by conditioning the model on D_1. This procedure is replicated over random splits of the data into D_1 and D_2. Of note, cross-validation automatically accounts for the dimensional penalty: a model that is overfitting (i.e. capturing non-reproducible random fluctuations) on D_1 is not expected to generalize well on unseen data D_2.

The entire procedure for cross-validation is computationally intensive and is thus not shown in detail here (see the manual available from the package for the detailed procedure). The results for the Microsporidia dataset are shown in Table 1. We see that the score of the CAT-GTR model (relative to LG) is much higher than that of CAT, which in turns has a better fit than GTR, and then finally LG. More generally, cross-validation on most empirical datasets over broad evolutionary scales (metazoans, eukaryotes, archaea, eubacteria, angiosperms, etc) invariably shows that site-heterogeneous models are much better fitting than one-matrix models, thus confirming the common-sense intuition that pattern heterogeneity is prevalent in empirical coding sequences.

Model	CV-score	standard deviation
GTR	288	28
CAT-F81	1154	122
CAT-GTR	2337	66

Table 1 Cross-validation scores (over 10 replicates) for the GTR, CAT-F81 and CAT-GTR models, relative to the LG model

2.5 Model checking by posterior predictive resampling

Measuring the empirical fit of alternative models represents a first fundamental principle for guiding the inference and the decision. On the other hand, it shows two main limitations. First, the fit of a model is a *global* measure of how well a model fits all aspects of the data, and not just those aspects that are relevant for the specific question being asked (here, the phylogenetic relationships). Thus, it can never be totally excluded that a lesser fitting model is in the end more accurate for phylogenetic reconstruction. Second, the fit is a *relative* measure, allowing one to determine the best among a series of models. Yet, even the best among the currently available models may not provide a good *absolute* fit to the data. For those reasons, it is essential to complement model comparison with model *checking* (i.e. using goodness-of-fit tests). Unlike model comparison, goodness-of-fit tests offer an absolute measure, by implementing a rejection test for each model taken individually. In addition, the test can be targeted, via the choice of summary statistics, to those aspects of the data that are deemed particularly relevant for the question being asked (Meng, 1994; Gelman et al., 2004).

In Bayesian inference, model checking is done using posterior predictive simulations. This method has been extensively used in Bayesian phylogenetics (Bollback, 2002; Lewis et al., 2014; Höhna et al., 2018). Posterior predictive checks can be seen as the Bayesian analogue of the parametric bootstrap: once the model has been conditioned on empirical data, the parameter configurations sampled from the posterior distribution are used to re-simulate replicates of the original dataset. Then, the value of some summary statistic of interest is computed on the simulated replicates, thus yielding a null distribution for the statistic under the fitted model. The value of the statistic computed on the original data is then compared to this null distribution. If it deviates significantly, this means that there is something in the empirical data which is not reproduced in the simulated replicates – and which is thus missed by the model.

The summary statistic used for the test should be designed so as to capture key features of the data that are deemed to be important in the context of the specific question under consideration. In the present case, we want to test each of the four models considered above, LG, GTR, CAT and CAT-GTR, concerning their ability to account for site-specific restrictions imposed by selection on amino acid usage. To this end, a simple statistic we can use is the amino acid *diversity*, i.e. the mean number of distinct amino acids per column across the sequence alignment. Under strong site-specific amino acid preferences, we expect a low diversity (a small subset of amino acids observed at each column).

Running a posterior predictive analysis can be done with the `readpb_mpi`, using the `-div` option for the diversity statistic (more general posterior predictive checks can be conducted using the `-ppred` option). In the case of the GTR model, the command is:

```
mpirun -np 8 readpb_mpi -x 500 1 -div gtrmicro1
```

and the program returns the following output:

```
diversity test
obs div : 4.07397
mean div: 4.67027 +/- 0.00923482
z-score : 64.5712
pp      : 0
```

The mean number of amino acids per column on the original alignment is 4.07. In contrast, the datasets simulated under the GTR model show on average 4.67 distinct amino acids per site. This difference is highly significant: the p-value is indistinguishable from 0, and the observed diversity is more than 64 standard deviations away from the mean of the posterior predictive null distribution (a p-value of 0.5 would approximately correspond to only 2 standard deviations away from the mean). In other words, the spectrum of amino acids present at each site in datasets simulated under GTR is too broad, compared to original sequence alignment, which indicates that the GTR model does not correctly reproduce (and thus, does not correctly capture) positional biochemical constraints. A similar result is obtained with LG (observed diversity is 72 standard deviations off the null distribution).

Doing the same experiment with CAT-F81 leads to a clearly different outcome:

```
diversity test
obs div : 4.07397
mean div: 4.07294 +/- 0.00810093
z-score : -0.126924
pp      : 0.5625
```

The diversity observed in data simulated under CAT-F81 is very close to the diversity of the original alignment (in fact, a bit smaller), and well within the posterior predictive null distribution ($p = 0.56$). This suggests that CAT-F81 adequately models site-specific amino acid propensities.

Finally, the CAT-GTR model, it is formally rejected by the test ($p < 0.01$):

```
diversity test
obs div : 4.07397
mean div: 4.0939 +/- 0.00850677
z-score : 2.3433
pp      : 0.00833333
```

However, the deviation between observed and posterior predictive diversity is much less than what was obtained above under the GTR or the LG model: the observed diversity (4.07) is now only 2.3 standard deviations away from the mean of the posterior predictive null distribution (4.09).

2.6 Pattern heterogeneity across sites and phylogenetic accuracy

We can now summarize the analysis and propose a global interpretation. Essentially two types of models were considered. On one side, LG and GTR assume pattern homogeneity across sites (i.e. invoke a single amino acid replacement process across all sites); on the other side, CAT and CAT-GTR explicitly account for site-specific amino acid preferences. Strikingly, the two models assuming pattern homogeneity give Microsporidia sister-group to all other eukaryotes, whereas the two models accounting for pattern heterogeneity give instead Microsporidia sister-group to Fungi.

Ignoring contextual knowledge about eukaryotic evolution, several independent lines of evidence suggest that site-heterogeneous models are more accurate in the present case. First, the tree produced by LG and GTR is a typical long-branch attract tree. Second, the much better relative fit of CAT-F81 and CAT-GTR, compared to LG and GTR, combined with the posterior predictive goodness-of-fit test using the diversity statistic, both show that site-specific amino acid preferences represent an important aspect of the true evolutionary process, which is not correctly captured by one-matrix models such as LG or GTR.

Why should the incorrect modeling of site-specific selective constraints make classical one-matrix models particularly sensitive to tree reconstruction errors? One main reason is that sites that are under strong biochemical constraints may nevertheless evolve rapidly – it is just that they stay within a small range of biochemically similar amino acids, which they keep repeatedly visiting. However, this fast evolution among a small number of possible amino acid states then makes it very likely for distantly related species to display the same amino acid at that site just by chance. Models ignoring site-specific amino acid preferences will underestimate this effect and will instead tend to incorrectly interpret the resulting identity by state as indicative of shared ancestry. As a result, they will underestimate sequence saturation and evolutionary distances (Halpern and Bruno, 1998) and will be more sensitive to long-branch attraction (Lartillot et al., 2007).

In the present case, we can get a rough estimate of the effective number of allowed amino acids per site by taking the exponential of the mean site entropy (8th. column of the trace file). Under the CAT-GTR model, this gives around 6.3 accepted amino acids per site, to be contrasted with 16.7 amino acids per site, according to the GTR model. Such a large discrepancy in the expectations under the two types of models as to the long-term probability of convergent evolution suggests that the effect of accounting for site-specific amino acid preferences on phylogenetic accuracy is likely to be substantial, and therefore represents a plausible explanation for the observed discrepancy between the two classes of models in the case of Microsporidia.

A similar phenomenon has been observed in several other well-characterized phylogenetic problems (e.g. Lartillot et al., 2007). More generally, there is now a broad array of empirical analyses showing that site-specific amino acid preferences represent an important aspect of the sequence evolutionary process, which is likely to negatively impact phylogenetic accuracy if not properly modeled. This is particularly true in the context of deep phylogenies (i.e. over broad evolutionary scales), for which sequence saturation is the rule and the risk of systematic errors in tree reconstruction is always an important concern. In the face of these problems, the site-heterogeneous models presented here certainly represent an important option to consider in a typical phylogenomic analysis.

In terms of practical recommendations, the best model is, by far, CAT-GTR, although the computational cost is high. A reasonable and computationally more efficient alternative is offered by CAT-F81. In spite of its rather crude approximations, CAT-F81 has generally proven more robust against long-branch attraction than one-matrix models in many situations (which is consistent with its good absolute fit under the posterior predictive diversity test). Therefore, a good procedure would be to always start with CAT-F81, so as to get a first idea of how much time it takes to obtain reasonable chains and to obtain a first series of useful results for the the dataset of interest. Then, if deemed affordable, a CAT-GTR analysis can be conducted (given that it would take about 6 to 8 times longer). A GTR (and possibly LG) analysis can also be conducted (possibly, with maximum likelihood implementations, such as RaxML [Stamatakis et al. 2005] or IQTree [Nguyen et al. 2015]). If the results are the same between site-homogeneous and site-heterogeneous models, then they can be considered

as robust. Conversely if the inferred tree topology turns out to depend on the model, then, a more thorough analysis along the lines proposed here can be conducted, using posterior predictive checks to make a stronger case about which model is likely to give a more accurate tree.

3 Challenges and perspectives

In summary, there is now good empirical evidence showing that accounting for pattern heterogeneity across sites makes an important difference when reconstructing deep phylogenies. The site-heterogeneous models implemented in PhyloBayes (CAT and CAT-GTR) currently represent the most radical approach – and perhaps the most accurate thus far – for capturing the modulations across sites in amino acid preferences. Over the recent years, the use of these models has proven instrumental for accurate inference on multiple practical cases.

All this comes at a cost, however. As it stands, the non-parametric random-effect models implemented in PhyloBayes are computationally very demanding. If the current implementation strikes a reasonable compromise between computational efficiency, model adequacy and phylogenetic accuracy for datasets of intermediate size (50 to 100 taxa, 10,000 to 30,000 positions), it does not scale up well to larger datasets, such as those that are currently contemplated in phylogenomics. For instance, resolving the relationships at the base of the metazoan tree, and addressing particularly difficult questions such as the position of ctenophores, seems to require datasets of the order of at 100,000 to 500,000 aligned positions (Simion et al., 2017). For such large datasets, the current implementation is not usable in practice, at least not directly on the full dataset.

In the face of these limitations, pragmatic solutions have been considered. A first simple approach is to use jackknife resampling, i.e. repeating the Bayesian analysis on subsets of genes or sites drawn without replacement from the original dataset and then averaging out the results over the replicates (Delsuc et al., 2008; Simion et al., 2017). Of note, this approach is not purely Bayesian and is admittedly a work-around. However, because of the additional layer of non-parametric resampling, jackknife should in fact produce robust estimates of the statistical support for the phylogeny.

Alternatively, other non strictly Bayesian approaches are currently being explored, most of which aim at proposing reasonable approximations explicitly accounting for site-specific amino acid preferences, while preserving the benefits of the increased robustness against tree reconstruction errors: posterior mean site frequency approximations (Wang et al., 2018), better empirical finite mixtures (Susko et al., 2018), penalized likelihood (Tamuri et al., 2014) or variational approaches (Dang and Kishino, 2019). These represent promising developments.

References

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., and Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387(6632):489–493.

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.*, 19(7):1171–1180.

Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., and Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.*, 54(5):743–757.

Dang, T. and Kishino, H. (2019). Stochastic Variational Inference for Bayesian Phylogenetics: A Case of CAT Model. *Mol. Biol. Evol.*, 36(4):825–833.

Delsuc, F., Tsagkogeorga, G., Lartillot, N., and Philippe, H. (2008). Additional molecular support for the new chordate phylogeny. *Genesis*, 46(11):592–604.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15(7):910–917.

Höhna, S., Coghill, L. M., Mount, G. G., Thomson, R. C., and Brown, J. M. (2018). P3: Phylogenetic Posterior Prediction in RevBayes. *Mol. Biol. Evol.*, 35(4):1028–1034.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc.*, 31:203–222.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *J. Am. Stat. Assoc.*, 90(430):773–795.

Koshi, J. M. and Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins*, 32(3):289–295.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.*, 7 Suppl 1:S4.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, 21(6):1095–1109.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, 62(4):611–615.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, 25(7):1307–1320.

Le, S. Q., Lartillot, N., and Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 363(1512):3965–3976.

Lewis, P. O., Xie, W., Chen, M.-H., Fan, Y., and Kuo, L. (2014). Posterior predictive Bayesian phylogenetic model selection. *Syst. Biol.*, 63(3):309–321.

Meng, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.*, pages 1142–1160.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32(1):268–274.

Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, 53(4):571–581.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.*, 9(3):e1000602.

Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.*, 22(5):1246–1253.

Quang, L. S., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.*, 67(5):901–904.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, E., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., and Manuel, M. (2017). A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol.*, 27(7):958–967.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463.

Susko, E., Lincker, L., and Roger, A. J. (2018). Accelerated Estimation of Frequency Classes in Site-Heterogeneous Profile Mixture Models. *Mol. Biol. Evol.*, 35(5):1266–1283.

Tamuri, A. U., Goldman, N., and Dos Reis, M. (2014). A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics*, 197(1):257–271.

Telford, M. J., Moroz, L. L., and Halanych, K. M. (2016). Evolution: A sisterly dispute. *Nature*, 529(7586):286–287.

Wang, H.-C., Li, K., Susko, E., and Roger, A. J. (2008). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.*, 8:331.

Wang, H.-C., Minh, B. Q., Susko, E., and Roger, A. J. (2018). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.*, 67(2):216–235.

Wang, H.-C., Susko, E., and Roger, A. J. (2014). An amino acid substitution-selection model adjusts residue fitness to improve phylogenetic estimation. *Mol. Biol. Evol.*, 31(4):779–792.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, 18(5):691–699.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA*, 87(12):4576–4579.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theor Biol*, 8(2):357–366.

Chapter 2.1 To What Extent Current Limits of Phylogenomics Can Be Overcome?

Paul Simion
Laboratoire d'Ecologie et Génétique Evolutive (LEGE), URBE
University of Namur, Namur, Belgium
polo.simion@gmail.com

Frédéric Delsuc[1]
Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE
Université de Montpellier, Montpellier, France

Hervé Philippe
Station d'Ecologie Théorique et Expérimentale, UMR CNRS 5321
Moulis, 09200, France
herve.philippe@sete.cnrs.fr

Abstract

Current phylogenomic methods are still a long way from implementing a realistic genome evolution model. An ideal approach would require a general joint analysis of genomic sequences, while including coding sequence annotation, protein evolution or gene transfer, among other mechanisms, to infer the complete evolutionary history of the studied genomes. Such an approach is computationally intractable and currently approximated by phylogenomic pipelines that implement a series of independent steps ranging from gene annotation to species tree inference or positive selection detection. Here we review the virtues and limits of current phylogenomic methods compared to what could be expected from an ideal method. We present five case studies to illustrate various issues and limits in current phylogenomic practices, while assessing their relative importance. We argue that data error is pervasive in modern datasets and models are still too simplistic compared to the complexity of biological and evolutionary processes. Importantly, joint analyses should be a research focus as the many steps of phylogenomic pipelines are not mutually independent. It is essential to recognize the hidden assumptions of the many types of analysis available to our community so as to circumvent model misspecifications and critically evaluate the relevance of their results. In conclusion, the quality of datasets should be enhanced via numerous, rigorous checkpoints, while also boosting the capability of models to handle biological complexity by the development of better models, particularly through joint analyses.

How to cite: Paul Simion, Frédéric Delsuc, and Hervé Philippe (2020). To What Extent Current Limits of Phylogenomics Can Be Overcome?. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 2.1, pp. 2.1:1–2.1:34. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Supplement Material https://github.com/psimion/SuppData_Simion_Chapter_2020_Limitations_Phylogenomics

[1] FD was funded by the European Research Council via the ERC-2015-CoG-683257 ConvergeAnt project.

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 2.1; pp. 2.1:1–2.1:34
A book completely handled by researchers.
No publisher has been paid.

1 Evolutionary history modeling and inference

Reconstructing genome evolution

The ultimate goal of phylogenomics is to reconstruct the evolutionary history of species through their genomes. In theory, this involves reconstructing the genomes of all organisms that ultimately supplied a DNA fragment to extant organisms and all events that generated modifications in the genetic material. Such details obviously cannot all be determined, but it is likely that several major patterns of great interest could be inferred, as illustrated in Figure 1. First, the evolutionary history of species generated through speciation and hybridization should have left a clear majority signal in genomes. Second, the signal left by horizontal gene transfers from distant or sister species (or so-called "gene flow") should be discordant from the majority signal. Third, the extent of incomplete lineage sorting would inform us about ancestral population sizes and times between successive speciation events. Fourth, mutations that became fixed because they provided a selective advantage could be differentiated from the bulk of neutral or slightly deleterious mutations (e.g. through an unexpected synonymous to non-synonymous substitutions ratio). Mutations can, for instance, involve single point changes, insertions of a few random nucleotides or a long stretch of nucleotides (from a transposable element or through illegitimate recombination), deletions of a few nucleotides or a long fragment (even complete chromosomes), duplications (of genomes or some chromosomes) or rearrangements (e.g. chromosome translocation, fission or fusion). The order of magnitude signals generated by each of these events may vary from a single point mutation to genome duplication. Hence, some could likely be finely characterized (e.g. the timing of a genome duplication) while it might only be possible to describe others statistically.

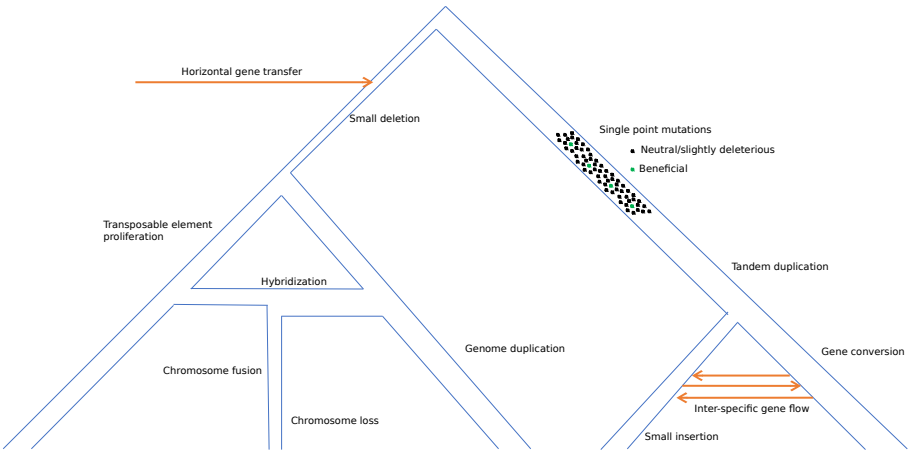


Figure 1 The historical processes shaping genome evolution. Schematic depiction of the main mutational processes that shape genomes. Their frequencies and impact on fitness are highly heterogeneous (e.g. from synonymous mutations to genome duplication or hybridization). These mutational processes are quite well known and relatively easy to model, whereas estimating the fitness of a given genome is much more difficult.

An ideal evolutionary model

The most natural way to infer this history from a series of genomes is to develop a genome evolution model and use standard statistical inference methods (in a Bayesian or maximum likelihood framework, see respectively Chapter 1.2 [Stamatakis and Kozlov 2020] and Chapter 1.4 [Lartillot 2020a]). The mechanisms depicted in Figure 1 have been the focus of massive in-depth studies for decades. Speciation is a pivotal theme in evolutionary biology, and DNA structure and change (including DNA repair) are crucial in molecular biology. All of this gives us an excellent idea of the most important mechanisms and how they work. So theoretically we have most of the knowledge required to develop a refined mechanistic model to reconstruct genome history. Naturally such a model can be designed in the mutation/selection framework (Bird, 1980), where the mutation process is independent of the DNA function, and a fitness function of the overall genome may be used to accept or reject a mutation. It is relatively easy to imagine how to model the mutational process inspired by simulators of genome evolution (Dalquen et al., 2012). For instance, single point mutations could be modelled with a general time reversible model (Tavare 1986; Chapter 1.1 [Pupko and Mayrose 2020]), and insertion/deletion with a hidden Markov model (Holmes and Bruno, 2001), while the mutational process would not necessarily be uniform across the genome (e.g. CpG hypermutability [Bird 1980] or proximity to the DNA minor groove [Pich et al. 2018]). Horizontal gene transfer should be considered as a mutation. Developing a fitness function is obviously much more difficult, but a function that only takes the major fitness components into account, i.e. non-coding RNAs and proteins, and their expression level, might be sufficient. A model similar to those used for gene annotation (see Chapter 4.1 [Necsulea 2020]) would enable prediction of non-coding RNA and protein sequences from genome sequences through the identification of transcription initiation sites and exon/intron structures. The fitness of these sequences could be estimated via a phenomenological approach (as in Yu and Thorne, 2006; Rodrigue et al., 2010). The expression level can be predicted based on promoter (nucleotide) and transcription factor (amino acid) sequences. Innovative solutions would certainly be required to be able to integrate all of these elementary fitness components into the fitness framework of a genome.

Ideal but beyond reach

Despite the attractiveness of such a theoretical model that could be used to infer major events which have occurred during genome evolution (see Figure 1), nobody has ever envisioned such a holistic approach. The reason may be that the approximations needed to compute genome fitness are so unrealistic that the extent of model violations would undoubtedly generate highly inconsistent results. But this is an unlikely explanation since, for instance, in phylogenetics the underlying maximum parsimony model, and to a lesser extent the Jukes-Cantor model are highly unrealistic, (e.g. based on the assumption that selective pressures are the same at every genome position). These models have nevertheless been and are still being frequently used. The most likely reason for not developing such a global genome evolution model is the tremendously high combinatorics. Sequence alignment and evolutionary tree inference independently constitute non-polynomial (NP) problems (see respectively Chapter 1.2 [Stamatakis and Kozlov 2020] and Chapter 2.2 [Ranwez and Chantret 2020]). As genomes are composed of millions of nucleotides, the number of possible ancestral genomes, evolutionary paths of organisms and DNA fragments is tremendous. Anyone who has ever tried to infer a phylogenetic tree from a relatively small dataset (e.g. 500 genes from 100 species) under the site-heterogeneous CAT-GTR model (PhyloBayes, Lartillot and

Philippe 2004) or a coalescent model (BEAST, Bouckaert et al. 2019) is aware of how far we are from making inferences with such a genome-scale model (see also Chapter 5.3 [Zhukova et al. 2020]).

Inferring the history of genomes therefore requires a divide and conquer approach combined with a clever choice of simplifying assumptions. The next section will roughly describe the main divisions that have been adopted by the phylogenomic research community.

2 The phylogenomic approach

Here we focus on species phylogeny inference using phylogenomics. We then exemplify the problems and advantages generated by the arbitrary division of a large-scale joint inference (see Figure 1) into several smaller elements (see Figure 2). Since less information (e.g. the evolutionary history of sequences is overlooked during alignment) is used at each step, errors may easily be made and their impact on subsequent steps is a concern. In contrast, working on a small-scale inference potentially allows us to use more complex models, hence reducing model violations.

2.1 A practical approximation

Divide and conquer

Essentially, reconstructing a given species phylogeny is a logical inference that uses both observable data (e.g. genomic sequences or morphological characters) and various premises (e.g. mutations are inherited through time, and transitions are more likely than transversions) to produce hypotheses regarding past evolutionary events. Unfortunately, the quantity and complexity of these premises currently hamper any practical holistic inference. With the aim of applying an approach capable of grasping the main aspects of the numerous evolutionary processes described in the previous section, while remaining practical, in phylogenomic analyses this integrated process is divided into independent blocks of computationally tractable units. These units form typical phylogenomic approaches, as illustrated in Figure 2, and they correspond to various common genomics procedures: (i) genome annotation, (ii) searching for homologous genes, (iii) defining orthologs, (iv) aligning homologous positions, (v) inferring species phylogeny (supermatrix or gene tree approaches), and (vi) reconciling single gene trees and the species phylogeny. Once the species tree is inferred, it is used in various methods to refine gene trees (duplication, loss, and horizontal gene transfer), infer the strength of selection, or reconstruct the gene order. Many of these procedures are detailed in other chapters of this book (see Chapters 1.4, 2.2, 2.4, 2.5, 3.2 and 4.5 [Lartillot 2020a; Ranwez and Chantret 2020; Fernández et al. 2020; Tannier et al. 2020; Boussau and Scornavacca 2020; Lowe and Rodrigue 2020]). The following description of conventional phylogenomic approaches is voluntarily brief and cursory and interested readers may find additional in-depth reviews on phylogenomics elsewhere (Delsuc et al., 2005; Philippe et al., 2005; Laumer, 2018).

Inferring orthologs

Phylogenomic pipelines are based on genomic data (e.g. often using coding sequences [CDSs] or ultra-conserved elements [UCEs]) and on transcriptomic data from multiple species. The choice of the molecular markers to be used is dependent on the biological question at hand, such as the evolutionary scale under investigation. These molecular datasets from multiple samples are then clustered into groups of homologous sequences. The criteria used

for homology prediction is pairwise sequence similarity estimated using BLAST (Altschul et al., 1990) or Smith–Waterman (Smith and Waterman, 1981) classical pairwise sequence alignment algorithms. Several algorithms have been proposed to make use of this pairwise similarity information to explicitly produce groups of homologous sequences, such as MCL (Enright et al., 2002), hcluster sg (Ruan et al., 2008) or SiLiX (Miele et al., 2011). Other tools use this pairwise similarity information to directly predict orthology relationships, such as Hieranoid (Kaduk and Sonnhammer, 2017), OrthoMCL (Li et al., 2003), OrthoFinder (Emms and Kelly, 2015), OMA (Altenhoff et al., 2019), Eggnog (Huerta-Cepas et al., 2019), UPhO (Ballesteros and Hormiga, 2016), Ortholog-Finder (Horiike et al., 2016) among others (see Chapter 2.4 [Fernández et al. 2020]), with variable complexity ranging from fairly straightforward (OrthoMCL) to complex procedures based on successive clustering loops and sequence alignments followed by gene tree inference and paralog splitting (Ortholog-Finder). Orthology prediction tools dovetail in many ways with the overall phylogenomic approach. They often rely on a variety of steps that include homology inference (i.e. similarity searches), pairwise species comparisons or species-overlap concepts, sequence alignment, gene genealogy inference, and even species tree inference or *a priori* knowledge of the species tree. Orthology prediction is thus often considered as a separate phylogenomic approach, and interested readers will find a more in-depth review of this topic in Chapter 2.4 (Fernández et al., 2020).

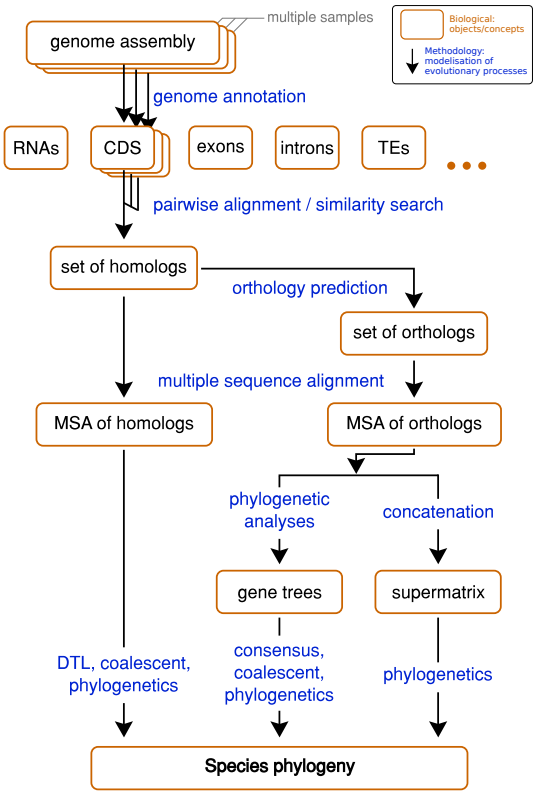


Figure 2 Common phylogenomic approaches. Schematic view of the series of practical analysis steps (in blue) of the phylogenomic approach.

Producing alignments

Multiple homologous or orthologous sequences are then jointly aligned using one of the many available sequence alignment software packages, such as MAFFT (Katoh and Standley, 2013), Clustal Omega (Sievers et al., 2011), MUSCLE (Edgar, 2004) or T-COFFEE (Notredame et al., 2000) (see also Chapter 2.2 [Ranwez and Chantret 2020]). Multiple sequence alignment software optimizes sequence alignment using various complex algorithms, but usually relative to a simplistic sequence evolution model (e.g. all substitutions and replacements are considered equiprobable) and the evolutionary history of the sequence is overlooked. Interestingly, one tool, i.e. MACSE v2 (Ranwez et al., 2018), models the codon structure of coding sequences during alignment. Using a more complex model unsurprisingly leads to better results and to some additional features, such as the ability to circumvent frameshifts that are often present in sequencing data (Ranwez et al. 2018; Chapter 2.3 [Ranwez and Delsuc 2020]).

Phylogenomic analyses

When using MSA of homologous sequences, directly concatenating them is impossible since multiple paralogous/xenologous sequences per species could be present. Instead, reconciliation methods are used to jointly analyse gene trees and species trees, notably by modeling gene duplication-transfer-loss (DTL) events. Available software includes Phyldog (Boussau et al., 2013) or ecceTERA (Jacox et al., 2016), among others (see Chapter 3.2 [Boussau and Scornavacca 2020]). Such analyses are complex but promising for the future of phylogenomics as they acknowledge the actual interdependency of two steps (i.e. jointly inferring gene and species trees), which are handled separately in the standard phylogenomic approach using orthologs only (see Figure 2). In addition, much more data can be used with these analyses than is possible with the reduced set of orthologs. Unfortunately, these methods are still very computationally-expensive to be widely used in a ML framework (Phyldog), although parsimony-based amalgamation methods such as ecceTERA could scale up with genomic data (ecceTERA).

Two main strategies are available to infer a species tree when using multiple sequence alignments (MSA) of orthologous sequences. Every alignment may be analysed independently to produce gene trees that may be incongruent because of incomplete lineage sorting (ILS), introgression or lateral gene transfer. This information may then be used to infer the underlying species tree, or otherwise every sequence per species may be concatenated in order to sum up their phylogenetic signals. There is ongoing debate on which strategy recovers the most accurate species trees (Springer and Gatesy, 2016; Edwards et al., 2016) and it is important to highlight three key arguments in this debate. First, taking ILS into account (see Chapter 3.4 [Bryant and Hahn 2020]) is impossible when using a concatenation approach which, despite the current use of more refined evolution models and more data, could never accurately solve a series of extremely fast speciation events given that it can be inconsistent under some evolutionary scenarios (Kubatko and Degnan, 2007). Second, single-gene tree reconstruction often yields little or no phylogenetic signal for difficult nodes (e.g. short internal branches) due to stochastic error. Third, only considering the species tree is not appropriate for subsequent evolutionary analyses (Hahn and Nakhleh, 2016). We believe that concatenation seems therefore more adequate to resolve ancient phylogenetic relationships or when the sampling is devoid of ultra-close speciation events, whereas the use of single gene trees is more appropriate for more recent speciation events, even when closely-spaced in time. Both methods rely on phylogenetic tree inference generally using

software based on ML, such as PHYML (Guindon et al., 2010), IQ-TREE (Nguyen et al., 2015), RAxML-NG (Kozlov et al. 2019; Chapter 1.3 [Kozlov and Stamatakis 2020]), or on Bayesian approaches such as MrBayes (Ronquist et al., 2012), BEAST (Bouckaert et al., 2019), and PhyloBayes (Lartillot et al. 2009; Chapter 1.5 [Lartillot 2020b]).

2.2 The costs of over-simplification and subdivision

While some model violations are often discussed concerning the phylogenetic inference step (e.g. ongoing debate on the development of the best sequence evolution models or on concatenation versus coalescence approaches), many other steps in the phylogenomic approach could also potentially lead to erroneous results. Below we discuss some of the problems encountered during the various practical steps in the phylogenomic approach when the ideal evolutionary model presented earlier is misspecified.

Information loss and implicit model violation

The subdivision of an ideal integrated model for reconstructing the evolutionary history of genomes into a series of independent blocks of computationally tractable units necessarily leads to the loss of potentially useful information, while forcing us to adopt an over-simplified model that cannot use the missing information. For instance, homology search through sequence similarity ignores the overall evolutionary history of the genomes being compared. Due to the loss of phylogenetic information, it implicitly makes the strong yet incorrect assumption that sequences were generated under a star-tree topology with equal branch lengths. However, the information that some species are closely related and that some others are fast-evolving is extremely useful for homology detection. The impacts of this model violation on the outcome are extremely hard to predict and study because of the substantial challenge of designing alternative non-star-tree models. Moreover, the model used to quantify similarity is extremely simplistic as it is solely based on an amino acid exchangeability matrix (e.g. JTT, or BLOSUM). It implicitly assumes that every position evolves at the same rate and that at most a single substitution has occurred at a given position, which are two obviously incorrect assumptions. This oversimplified model explains the poor sensitivity of the BLAST score (Koski and Golding, 2001). Interestingly, alongside the publication of the orthology inference tool Orthofinder (Emms and Kelly, 2015), the authors designed a blast score double-normalization. It normalises BLAST scores for alignment length and, more importantly, these pairwise scores are normalised across species according to their evolutionary distances, so it is striving to transform the scores as if the sequences had been generated under a star-tree topology. This interesting approach nevertheless cannot control saturation of the similarity score, which means that the correction will be much more accurate for closely related species than for divergent ones.

Genome annotation errors

As briefly introduced in Figure 2, genome annotation is one of the first steps of most phylogenomic pipelines. Annotating genes requires a set of complex methods that rely on knowledge regarding genetic code, intron structure, transcription and translation mechanisms or RNA-seq data (see Chapter 4.1 [Necsulea 2020]). Yet, it often assumes that genomes do not have any evolutionary history, again an obviously false assumption. A shortcut to input some evolutionary information is to compare predicted coding sequences with transcriptomes or proteomes from closely related species (Dunne and Kelly, 2017; Monnahan et al., 2019; Rey

et al., 2019). Unfortunately, current annotation methods do not model chromosome structure, protein folding or interaction with other genomic regions. These limitations lead to erroneous gene predictions that can ultimately mislead comparative genomic analyses (see examples in Section 3.1).

Sequence alignment model violations

Multiple sequence alignments are also hampered by model violations (see Chapter 2.2 [Ranwez and Chantret 2020]), i.e. some mutational processes are explicitly modelled while overlooking sequence function and protein structure as well as their chromosome-wise context, such as species-specific recombination hotspots or even lineage-specific evolutionary rates (i.e. heterotachy). Otherwise, when aligning multiple sequences, indels are implicitly considered as characters rather than historical events. The latter is a misspecification of an ideal evolutionary model, which is tackled by the dynamic homology concept. This issue has led various authors to develop methods for joint inference of sequence alignments and species trees (Fleissner et al., 2005; Redelings and Suchard, 2005; Herman et al., 2014; Wheeler et al., 2015). As expected, this interesting approach is computationally intensive, thus seriously limiting the dataset size and the complexity of the sequence evolution model that can be handled.

Unrealistic phylogenomic inference models

In contrast, phylogenomic analyses of aligned and concatenated sequences enable the use of more complex evolutionary models geared towards minimizing model violations. However, some potentially important aspects of genome evolution are still not taken into account by most phylogenomic inference methods, e.g. lineage-specific composition heterogeneity, site-specific substitution process heterogeneity, or heterogeneity of site-specific substitution process among lineages (i.e. heteropecilly, Roure and Philippe 2011). Note that even when some methods are available to model one aspect of genome evolutionary processes, e.g. modeling ILS, site-heterogeneity or DTL in reconciliation methods, it is seldom feasible to combine them, and if it were, the resulting implementation would surely be extremely time-consuming. For example, combining a CAT model with a GTR component, a Gamma component, amino acid compositional breakpoints along the tree and evolutionary rate breakpoints along the tree while allowing for gene transfer across lineages to analyse relationships between 300 complete genomes would clearly be beyond reach with current computation resources. Finally, knowledge on the genomic context of a sequence is still not used in the phylogenetic inference process.

Software errors

In addition to these errors — for which we know the origin albeit we do not know where they are in the dataset — there are unknown errors, i.e. errors in the implementation such that the script/software does not produce the intended results. These unknown errors are expected because limited funding and publish-or-perish pressure imply that an insufficient amount of time is generally devoted to quality control of both programs (Czech et al., 2017; Darriba et al., 2018) and pipelines (see Section 4.5).

All of the severe model violations described above, albeit unavoidable for computational tractability reasons, as well as information loss very likely generate errors at each phylogenomic pipeline step. Importantly, of all these errors accumulate along the pipelines, with a possible snowball effect. For instance, annotation errors alone will generate additional errors

in homology detection, which in turn will generate more errors in the alignment and finally in phylogenetic inference. Hereafter we briefly discuss the robustness of phylogenomics to these errors and how to reduce their impacts.

3 Relative robustness to pervasive errors

3.1 Types of error and methods to detect them

Theoretical limitations of the successive independent and simplistic steps of the phylogenomic approach inevitably lead to the production of errors. These can appear at all steps of a given pipeline and can propagate from step to step. Here we briefly describe various error types and some recent methods or tools that can detect them and thus reduce their impact on the phylogenomic approach as a whole. We have classified these errors into three arbitrary groups: i) observational errors during data acquisition and production, ii) errors during dataset assembly, and iii) errors during phylogenetic inference. These errors can be generated by experimental error (e.g. contamination by DNA from other species), stochastic error (e.g. insufficient coverage), and systematic error (i.e. due to model violations).

Observational errors

This type of error concerns data that are not what the user believes they are. These include contamination from organisms other than the target (e.g. bacteria, fungi, trypanosomes, viruses), cross-contamination between samples during sequencing data production, sequencing and assembly errors, fragmented transcriptomic contigs thought to be entire transcripts, gene exons thought to correspond to entire genes, gene introns thought to correspond to exons, amino acid sequences translated out of frame (i.e. frameshifts). Contamination in genomic data can partially be detected by Blobtools by combining coverage, GC content, and blast taxonomy (Laetsch and Blaxter, 2017), large scale similarity search with Conterminator (Steinegger and Salzberg, 2020) or by the consensus of various methods (Cornet et al., 2018). Contamination is not only present in the data generated during a given study, but also affects public databases: e.g. 5% of the publicly available cyanobacterial genomes turned out to be highly contaminated (Cornet et al., 2018) and a recent large-scale analysis of GenBank identified more than 2,000,000 contaminated sequences! (Steinegger and Salzberg, 2020). Cross-contamination affects both DNA and RNA data and is increasingly acknowledged as a pervasive issue (Ballenghien et al., 2017; Alié et al., 2018; Allio et al., 2020; Prous et al., 2020). It can be handled by the CroCo program which relies on coverage to detect the actual origin of a sequence in a set of samples (Simion et al., 2018). It has been shown that up to 30% of transcripts from a *de novo* assembled transcriptome could be cross contaminated (i.e. actually belong to another species, Simion et al. 2018) or up to 26% of ddRAD loci (Prous et al., 2020).

Assembly errors during *de novo* transcriptome assembly (e.g. fragmentation due to insufficient coverage) can be corrected by fusing non-overlapping fragmented transcripts based on a multi-species orthology context, as shown in Section 4.1, where 30.6% of the transcripts were fragmented (124,096 out of 405,055 transcripts analysed). Annotation errors are also pervasive, as illustrated by the fact that reannotation of the well-known model group *Drosophila* recently led to the discovery of 500 to 1,000 new genes per species (Yang et al., 2018). Recent studies have proposed tools to correct gene annotation based on comparisons with other species (Dunne and Kelly, 2017; Rey et al., 2019; Monnahan et al., 2019). Using a non-overlapping sequence criteria on gene annotation from tunicate genomes (see Section

4.1), we estimate that 5.6% of the predicted genes were split into, usually, two exons (3,178 out of the 56,694 genes analysed). A next step towards more holistic genome annotation approaches could involve joint annotation of several genomes at once, with tractable computations if the species tree is known. Current tools using such a comparative framework unfortunately only aim at correcting gene prediction *a posteriori* (Dunne and Kelly, 2017; Monnahan et al., 2019) or at improving transcriptomic assemblies that could then be used to help gene prediction (Rey et al., 2019). Finally, frameshifts (due to sequencing, assembly or annotation errors) produce very divergent sequence stretches. These can be corrected during sequence alignment by taking the codon structure of coding sequences into account using the MACSE v2 alignment tool (Ranwez et al., 2018). If already present, frameshifts can be detected and masked in a multi-sample alignment using two recent segment filtering tools, i.e. HMMcleaner (Di Franco et al., 2019) and Prequal (Whelan et al., 2018).

Orthology errors

Similarity searches, usually via BLAST (Altschul et al., 1990), are hampered by bias, where higher scores are given to long dissimilar alignments than to short highly similar ones. This bias, combined with the lack of precision of the method in detecting very dissimilar sequences, means that a poorly assembled or highly divergent transcriptome will likely yield poor homology results. Pairs of homologous sequences have different potential relationships: they can be orthologous (i.e. stemming from speciation), paralogous (i.e. stem from duplication) or xenologous (i.e. stemming from horizontal transfer) — see Chapter 2.4 (Fernández et al. 2020) for an in-depth review of this topic. When using orthology inference algorithms, sequence pairs can be erroneously categorised as a subtype due to several issues. For example, incomplete taxonomic sampling complicates the detection of xenologs Kuzniar et al. (2008), and inaccurate gene tree inference can hampers the classification of two sequences as orthologous (e.g. if only one of them evolved rapidly). While ongoing research is focused on improving orthology inference tools, it is also crucial to design a taxonomic sampling method tailored to the evolutionary scale at hand to ensure accurate inference of sequence relationships. Orthology sets can be further improved by *a posteriori* checking orthology set consistency (Simion et al., 2017). Lastly, several tools can eventually be used to analyse sequence alignments and gene trees in order to detect and limit the impact of orthology errors on the phylogenomic pipeline, e.g. Phylo-MCOA (de Vienne et al., 2012), treespex (Struck, 2014), Branch Length Correlation (BLC) methods (as in Simion et al. 2017), Treeshrink (Mai and Mirarab, 2018) or reconciliation methods (Dondi et al., 2016). See a comparison of some of these methods in Section 4.2.

Evaluating phylogenomic datasets

It should be noticed that phylogenomic pipelines could be highly diversified in terms of implementation, thus leading to highly diversified phylogenomic datasets. Unfortunately, these datasets are seldom compared with statistics other than simply the numbers of genes and species included. Looking beyond summary statistics in these phylogenomic datasets can reveal various levels of data quality (e.g. measured with the Robinson-Foulds distance between gene trees and species trees), with orders of magnitude difference in data quantity (see Figure 2 in Simion et al. 2017, and Figure S4D in Philippe et al. 2019). For instance, regarding the debated phylogenetic position of ctenophores, less than 30% of the gene tree bipartitions are congruent with the species tree in three datasets that support the ctenophora-first hypothesis (Dunn et al., 2008; Hejnol Andreas et al., 2009; Moroz et al., 2014), whereas more

than 60% are congruent in a dataset that supports the porifera-first hypothesis of Simion et al. (2017). Data quality governs the crucial phylogenetic signal-to-noise ratio upon which the accuracy of the inferred species tree strongly depends. When working on debated species phylogeny, we stress the need to carefully inspect the phylogenomic pipelines used and the respective virtues of the datasets they led to, as their signal-to-noise ratio might be pivotal in evaluating the reliability of a given phylogenomic result. See also Chapter 2.5 (Tannier et al. 2020) for an original approach to gene tree quality assessment based on ancestral genome reconstruction.

Phylogenetic inference errors

Many sequence evolution models are available for phylogenetic inference based on homologous sequence alignments. Early models were tailored for single-gene analyses but the datasets have increased in both their dimensions (i.e. more markers and taxa) and complexity. Model assumptions are now recognised as often being violated by complex datasets, thus prompting the need to also increase the model complexity. For example, introduction of the Gamma component in models discredited the assumption that all sites evolve at the same pace (Yang, 1994), and more recent site-heterogeneous CAT models refuted the assumption that all sites evolve under the same substitution process (Lartillot and Philippe 2004 and Chapter 1.4 [Lartillot 2020a]). Potential model misspecifications are still numerous and could result in erroneous topology inference (e.g. LBA artefact, compositional bias) and incorrect branch length estimation. Various methods and models have been developed to reduce these misspecifications, such as GHOST models implemented in IQ-TREE in order to model heterotachy (Crotty et al., 2019), CAT models to phenomenologically account for protein structure and function (Lartillot and Philippe, 2004), the PMSF approach recently implemented in the ML framework (Wang et al., 2018), compositional breakpoints (BP) to account for heterogeneity in the substitution process across lineages (Blanquart and Lartillot, 2006, 2008) or site-heterogeneous codon models (SelAC) to model stabilising selection (Beaulieu et al., 2019). Data recoding has a special role in current phylogenomics. It consists in grouping different character states into a single common character leading to alphabet reduction (e.g. the Dayhoff 6-state recoding scheme). It is used in order to reduce compositional bias and saturation in the data, thus enhancing the phylogenetic signal (Susko and Roger, 2007). The relative importance of signal loss in comparison with the reduction in compositional bias and saturation is still debatable and likely depends on the characteristics of the dataset under study. In our opinion, data recoding should be considered suitable for large supermatrices only and recoded datasets are still highly complex so that they still require to be analysed with complex models (e.g. CAT models, see Feuda et al. 2017). Finally, a recent study assessed the impact of modelling site-heterogeneity versus partition-wide heterotachy and convincingly concluded that modelling site-heterogeneity was more important than modelling partition-wide heterotachy (Wang et al., 2019).

So far we have discussed ways to limit errors stemming from systematic bias, but when the phylogenetic signal is weak, stochastic errors can occur even when using large phylogenomic datasets. This is particularly true for ancient and short internal branches. In fact, with such difficult relationships, the data quantity required is so high that even a large sampling of complete genomes would not be enough to resolve them (Philippe et al., 1994).

3.2 Consistent species phylogenies

A large quantity of signal

The phylogenomic approach is, despite its flaws, surprisingly robust, as most pipelines will lead to the recovery of a similar species tree topology. This can be explained by the sheer quantity of phylogenetic signal accumulated when thousands of molecular markers are combined. This is not surprising, as many parts of the tree of life have already been correctly inferred using comparatively small morphological character matrices or single gene phylogenies. Phylogenetic signal is additive, so the amount of signal increases with the data quantity. In fact, only additive errors can compete with phylogenetic signal by producing a non-phylogenetic signal, leading to the recovery of an erroneous tree. For an error to be additive it has to produce the same kind of bias repeatedly across markers and lineages. For example, genomic cross-contamination from sample A to sample B will repeat the same mislabelling of B sequences, and species B will eventually be attracted towards the phylogenetic position of species A (Laurin-Lemay et al., 2012). As another example, in a lineage that has a naturally high evolutionary rate, on average all markers will present more homoplasy with another fast-evolving lineage, and both long-branch lineages will attract each other (i.e. LBA artifact). Conversely, various randomly distributed errors will only produce non-additive signals (often called "noise') that will not severely distort the phylogenetic signal. Even if not correctly modelled, these errors will simply reduce the statistical power of phylogenomics (see Section 4.3). Overall, the phylogenomic approach produces a globally consistent species tree as long as the phylogenetic signal prevails over the systematic error.

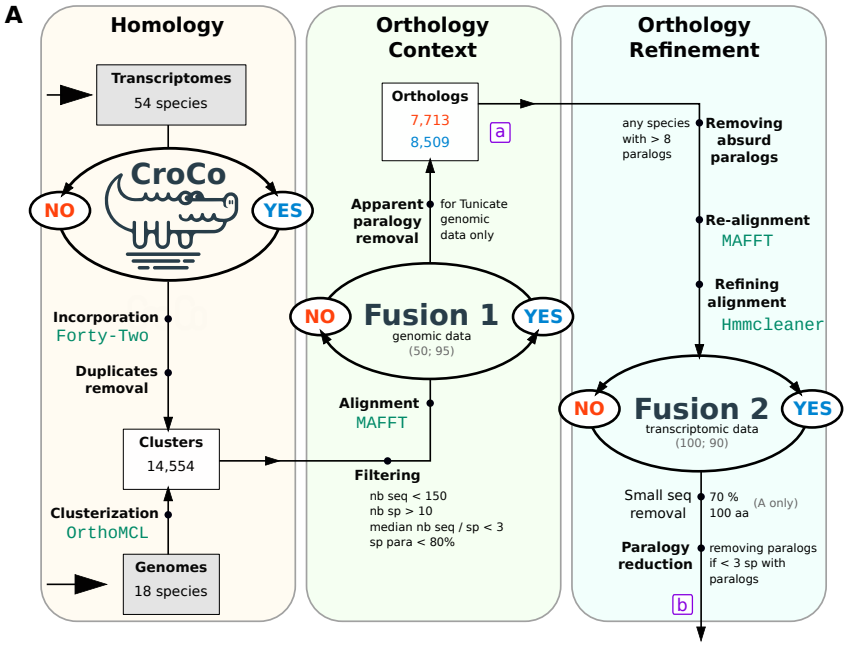
Few very difficult cases

Phylogenomic inconsistency only occurs in a few cases across the tree of life, all of which share the same characteristics and correspond to short internal branches. These branches bear a limited amount of phylogenetic signal so they are highly susceptible to errors, even random errors (e.g. see Section 4.3). Indeed, the signal-to-noise ratio for these branches is so low that any perturbation or noise will hamper signal extraction regardless of the intrinsic qualities of the model used. Branches are short when diversification has occurred rapidly through time, and the problem becomes more complex when the speciation event was too ancient, with progressive loss of the historical signal through multiple substitutions (i.e. saturation). Difficult relationships stemming from the first case triggered ILS modelling research (see Chapter 3.3 [Rannala et al. 2020]), while relationships derived from the second case underscore the need for better models to optimize the efficiency of extraction of the scant amount of remaining historical signal (see above and Chapter 1.4 [Lartillot 2020a]). As the phylogenomic approach is largely consistent across species phylogenies, except for short internal branches with low signal-to-noise ratio, it is not surprising that long-standing phylogenomic disputes are finally now focused on a few difficult relationships, i.e. the phylogenetic position of ctenophores, xenacoelomorphs, Stauromedusae, Laurasiatheria, the root of the placental tree or the early evolution of birds and eukaryotes.

4 Case-studies: Examples of current limits of phylogenomics

4.1 Correcting data errors in tunicates

Current phylogenomic practices involve similar handling of genomic and transcriptomic data from orthology inference to final sequence alignment, which violates the hypothetical ideal



B

	Orthologs at checkpoint a	Orthologs at checkpoint b	Min species recquired	Final Orthologs	congruence gene tree/ species tree	Max Missing% per site	Supermatrix length	Global Missing%	Total known residues
Standard Pipeline	**7,713**	**2,416**	5	2,024	mean = **63.3 %**	100	156,373	65.3	3,852,366
			10	1,770		80	95,037	46.9	3,582,646
			20	801		**50**	**54,878**	**34.9**	**2,536,964**
			30	**408** →		20	3,139	17.1	184,735
Corrected Pipeline	**8,509**	**6,732**	5	5,341	mean = **67.2 %**	100	1,860,030	57.5	56,067,810
			10	5,161		80	1,267,890	40.8	53,330,401
			20	3,993		**50**	**849,403**	**29.3**	**42,614,055**
			30	**3,039** →		20	210,163	16.0	12,534,696

Figure 3 Impact of data correction on tunicate phylogenomics. A) Detailed procedure for both standard and corrected pipelines. Three steps were only used in the "corrected" procedure: CroCo, Fusion 1 and Fusion 2. None of these steps were used in the "standard" pipeline. B) Summary statistics of the datasets produced from the two pipelines. The left side of the table corresponds to statistics obtained from the final orthologs highlighted in bold (i.e. 408 and 3,039). Conventional criteria for missing data filtering in supermatrices are highlighted in bold for both pipelines to ease comparison for readers.

holistic evolution model described at the beginning of this chapter. Indeed, a mix of genomic and transcriptomic data is often used in phylogenomic studies as transcriptomic data sequencing and assembly is both cheaper and faster than whole genomes. Transcriptomics is a cost-efficient way to enrich taxonomic sampling, which in turn helps to infer more accurate trees. Both types of sequencing data are potentially subject to contamination and cross-contamination. However, genomic and transcriptomic data differ in their nature. One the one hand, genomic data quality relies on accurate gene annotation, genes can occur in multiple copies and it might be expected that a genomic gene set is exhaustive for a given organism. On the other hand transcriptomic data quality mostly relies on the transcriptome assembly accuracy, whereas alternative splicing is often overlooked in favour of keeping the longest transcript, and transcriptomes are not exhaustive as rare transcripts are likely to be missed if the sequencing coverage is not deep enough. Regardless, both types of data are usually handled the same way. Data errors must be properly modelled in order to reduce discrepancies between data acquired by current protocols and real biological data. Here,

we simply tried to *a posteriori* correct some misspecifications, namely cross-contamination, split gene annotation and fragmentation of transcriptomic assemblies. We assessed the importance of these misspecifications by comparing datasets built with two different pipelines on the same mix of tunicate genomic and transcriptomic assemblies.

Both pipelines start with a set of putative orthogroups created with 18 genomes using OrthoMCL and then follow a series of filters and alignments until a final set of orthogroups is reached. The first pipeline is straightforward as it does not try to correct potential errors impacting genomic and transcriptomic data (no CroCo, no Fusion 1 and no fusion 2, see the *standard* pipeline in Figure 3A). The general procedure is designed as follows: transcriptomic data are incorporated into clusters of homologous sequences based on genomic data using Forty-Two (see https://bitbucket.org/dbaurain/42) and only alignments with reasonable size and diversity are kept. We then check tunicate species (only those based on genomic data) and remove all alignments in which at least one species presents two sequences (i.e. apparent paralogy). For each alignment, we then remove all sequences from species (based on transcriptomic data) that have too many paralogous sequences (i.e. more than eight copies per species). Short sequences that span less than 100 amino acids and have more than 70% missing data in a given alignment are discarded. Lastly, if less than three species present paralogs for a given alignment, their sequences are removed. The second pipeline adds three steps to the previous one (CroCo, Fusion 1 and Fusion 2, see the three blue "YES" of the *corrected* pipeline in Figure 3A). First, transcriptome assemblies from 54 species are cleaned from cross-contaminations using CroCo (Simion et al., 2018). Second, gene annotations of tunicate species (only those based on genomic data) are refined in the comparative context of an alignment by fusing together several sequences from the same species when they do not overlap. This is the "Fusion 1" step. Third, tunicate fragmented transcripts (based on transcriptomic data) are improved by also fusing same-species non-overlapping sequences in the alignments. This last "Fusion 2" step was only possible after an orthology context was reached by filtering out alignments containing paralogy for tunicate genomic data thus ensuring that we did not erroneous fuse non-overlaping paralogs.

The simultaneous use of the three corrections described above combined with paralogy filters led to a dramatic increase of roughly one order of magnitude in the size of the assembled dataset (see Figure 3B). The corrections improved the quantitative metrics measured here, i.e. gene number (408 to 3,039) and missing data (34.9 to 29.3%), thus increasing the number of known residues by 17-fold (2.5 M to 42.6 M). Gene alignments and supermatrices are available on the following website, containing the Supplementary Material of the current article: `https://github.com/psimion/SuppData_Simion_Chapter_2020_Limitations_Phylogenomics` . A recent study also reported a marked improvement in their dataset after cross-contamination removal with CroCo, with a gene number increase of 2,993 to 6,621 (Allio et al., 2020). This improvement was associated with an impact on the species tree topology and branch lengths, as expected given the findings of previous studies on the impact of the presence of cross-contamination in phylogenomics (Laurin-Lemay et al., 2012; Simion et al., 2018)). Importantly, these quantitative improvements were observed hand in hand with a qualitative improvement of the phylogenomic dataset. Indeed, the congruence between gene trees and species tree increased from 63.3% to 67.2% when using our three correcting steps (Figure 3B).

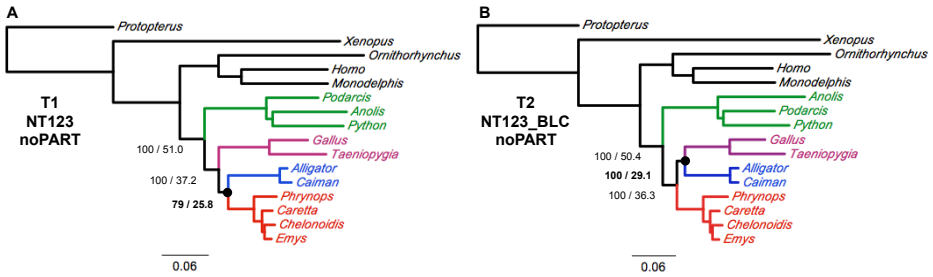
Why do cross-contamination removal and non-overlapping sequence fusion lead to a dataset that is an order of magnitude larger and of better quality? The answer lies in the existing interdependency between the different steps of the phylogenomic approach. We handled each step independently, but they share many assumptions that underlie an ideal genome evolu-

tion model, which means that a misspecification in one step will impact the next one. First, orthology inference tools are based on the assumption that sequences are correctly associated with an organism (i.e. no cross-contamination). Second, the accuracy of filtering clusters corresponding to 1-to-1 orthologs depends on the extent to which the genes are complete (i.e. correct gene annotation). In our example, gene annotation was *a posteriori* improved by being considered in the comparative context of a multi-species alignment. Third, high transcriptomic assembly quality (i.e. no fragmented transcripts) is essential when considering contigs from a transcriptomic assembly as biological transcripts. By carrying out simple *a posteriori* corrections to some known issues in the practical phylogenomic approach, we were able to better take into account some evolutionary and experimental processes that produced the genomic data under study. Orthology inference methods are more accurate if contaminants are removed and the number of apparent paralogs is reduced if split gene annotation and transcripts are fused back together. These simple corrections ultimately led to a vastly larger phylogenomic dataset.

Although useful, our simple corrections are still insufficient to build a truly genome-scale dataset. Indeed, a large percentage of the genes in the genomes and transcriptomes are still not present in the supermatrices. This might be due to incomplete sequencing and assembly of genomes and transcriptomes, to the limits of the orthology assignment method used (e.g. missing small and fast evolving genes) and/or to incomplete taxonomic sampling, which hampers accurate reconstruction of complex evolutionary histories (e.g. transfers, duplications, losses).

4.2 Cleaning outlier sequences and genes in turtle phylogenomics

The phylogenetic position of turtles within amniotes offers a great example of a longstanding question that has finally been answered through the resolving power of the phylogenomic approach. In 2012, two phylogenomic studies based respectively on transcriptomes (Chiari et al., 2012) and ultra-conserved DNA elements (Crawford Nicholas G. et al., 2012) independently found convincing support for positioning turtles as a sister group of archosaurs (birds and crocodiles), to the exclusion of lepidosaurs (lizards and snakes). This more derived position of turtles, recently confirmed by a larger scale phylogenomic analysis (Irisarri et al., 2017), implies that the anapsid condition of turtles (no temporal fenestration) is a derived state, whereas it was classically interpreted as the ancestral condition for amniotes. Chiari et al. (2012) built a phylogenomic dataset of 248 single copy nuclear genes for 16 vertebrate taxa, which was assembled according to best reciprocal hits obtained with BLAST based on the genomes and orthology annotations available at the time, along with newly generated transcriptomes. They showed that ML and Bayesian concatenations, and gene trees/species tree approaches performed under the best fitting nucleotide and amino acid substitution models unambiguously supported the classification of turtles as a sister group to birds and crocodiles (T2 in Figure 4B). However, the use of more simplistic nucleotide substitution models for both concatenation and gene trees/species tree reconstruction methods led to an alternative topology by artifactually grouping turtles and crocodiles (T1 in Figure 4A), likely because of third codon position saturation. This 248-gene dataset has since become an exemplary dataset in several studies aimed at testing phylogenetic reconstruction methods by comparing concatenation versus gene trees/species tree approaches (Bayzid et al., 2014; Mirarab et al., 2014, 2016; Simmons et al., 2016, 2019; Gatesy et al., 2019). Moreover, two recent studies have used this dataset as a core example to illustrate methods to detect outlier genes in phylogenomic analyses based respectively on Bayes factors (Brown and Thomson, 2017) and gene-wise likelihoods (Walker et al., 2018) between alternative



C

	AA	AA_BT	NT12	NT12_BT	NT12_TS	NT12_BLC	NT123	NT123_BT	NT123_TS	NT123_BLC
Genes	248	246	248	246	248	237	248	246	248	237
Sites	62,342	62,011	124,684	124,022	124,626	120,054	187,026	186,033	186,939	180,081
Variable sites	21,172 (33.96%)	21,034 (33.92%)	32,556 (26.11%)	32,352 (26.09%)	36,167 (29.02%)	36,282 (30.22%)	87,388 (46.73%)	86,868 (46.69%)	90,528 (48.43%)	89,311 (49.60%)
Missing	35.25%	35.26%	35.25%	35.26%	37.35%	35.81%	35.25%	35.26%	37.35%	35.81%
noPART BP (T1)	-	-	-	-	-	-	79	-	73	-
noPART BP (T2)	100	100	100	100	100	100	-	97	-	100
noPART CF_{gene} (T1)	-	-	-	-	-	-	25.8	-	28.3	-
noPART CF_{gene} (T2)	28.6	28.5	26.6	26.0	25.8	27.4	-	28.0	-	29.1
PART BP (T1)	-	-	-	-	-	-	-	-	-	-
PART BP (T2)	100	100	100	100	100	100	78	100	67	100
PART CF_{gene} (T1)	-	-	-	-	-	-	-	-	-	-
PART CF_{gene} (T2)	30.6	29.3	26.2	26.8	25.8	27.8	29.4	29.3	29.9	31.6

Figure 4 Assessing the effects of three filtering methods and sequence evolution models on the phylogenetic position of turtles. A) Maximum likelihood topology (T1) obtained with IQ-TREE under a single concatenated GTR+G model on the original nucleotide dataset of Chiari et al. (2012), including all codon positions. B) Maximum likelihood topology (T2) obtained with IQ-TREE under a single concatenated GTR+G model on the nucleotide dataset, including all codon positions filtered using the branch length correlation (BLC) method of Simion et al. (2017). Numbers at nodes indicate the standard ML bootstrap percentage/gene concordance factor, respectively. Bullets indicate nodes for which support values are reported in the table. Scale is based on the mean number of substitutions per site. C) Summary statistics and support values for the alternative T1 and T2 topologies obtained with different datasets (AA, NT12, and NT123) resulting from three following filtering methods: BT, TS, and BLC. The datasets were analysed under a single concatenated model (noPART) and a partitioned model by gene (PART). Abbreviations: BT: removal of the two paralogous genes identified by Brown and Thomson (2017); TS: TreeShrink filtering with default parameters (Mai and Mirarab, 2018); BLC: filtering with the branch length correlation method of Simion et al. (2017); AA: amino acid dataset; NT12: nucleotide dataset with saturated third codon positions removed; NT123: nucleotide dataset with all codon positions; noPART: single concatenated model (LG+G for amino acids and GTR+G for nucleotides); PART: partitioned model by gene (best models determined with ModelFinder); BP: ML bootstrap percentage; CFgene: gene concordance factor.

topologies. Brown and Thomson (2017) detected two genes as marked outliers within the 248 single-copy orthologous genes of Chiari et al. (2012) and showed that the corresponding alignments contained non-orthologous sequences, thus creating conflicting gene trees and resulting in the artefactual topology (T1) grouping of turtles and crocodiles when analysing the complete concatenated nucleotide dataset.

Protocols have been proposed to clean outlier sequences from alignments based on branch length analysis of the corresponding gene trees. Analyses of branch lengths between concatenation trees and those of individual gene trees have been used to exclude genes in phylogenomic analyses of metazoans (Simion et al., 2017), as well as to curate single-copy

orthologous gene alignments of the OrthoMaM database (Scornavacca et al., 2019), and to exclude outlier sequences when focusing on terminal branch lengths (Simion et al., 2017). This approach is here referred to as Branch Length Comparison (BLC). A similar method to detect outlier sequences that artificially inflate the diameter of individual gene trees was recently developed and implemented in the TreeShrink software package (Mai and Mirarab, 2018). Here we used the dataset of Chiari et al. (2012) to illustrate the impact of different cleaning methods on resolving phylogenetic conflicts regarding the position of turtles. First, we removed the two paralogs identified by Brown and Thomson (2017) of the original amino acid and nucleotide datasets (BT). Second, we used HMMCleaner to remove likely non-homologous sequence fragments from the original nucleotide datasets, removed residual sequences shorter than 50 nucleotides, inferred individual ML gene trees and concatenated trees under a GTR+G model using RAxML 8 (Stamatakis, 2014). Third, we applied TreeShrink with default parameters (TS, Mai and Mirarab 2018), and used the method of Simion et al. 2017 –here named BLC– to exclude outlier sequences having a terminal branch-length ratio >5 and gene alignments with an R2 Pearson correlation coefficient between all branch lengths in each gene tree and the corresponding supermatrix tree outside of the normal distribution (mean +/- 1.96 standard deviation). Finally, we performed phylogenetic reconstruction, along with gene and site concordance factors, on the resulting datasets using IQ-TREE (Minh et al., 2018) under a single concatenated LG+G or GTR+G model (noPART) and a gene partitioned model (PART), with model selection performed using ModelFinder (Kalyaanamoorthy et al., 2017) on amino acid datasets (AA), nucleotide datasets with only first and second codon positions (NT12), and nucleotide datasets with all codon positions (NT123).

The application of TreeShrink (TS) resulted in the removal of a total of 82 sequences in 76 gene alignments, whereas the BLC method removed 10 outlier sequences and 11 genes. Only BLC allowed automatic detection and exclusion of the two alignments containing paralogous sequences (ENSGALG00000008916 and ENSGALG00000011434). TS only excluded the *Monodelphis* sequence from ENSGALG00000008916. As previously shown by Brown and Thomson (2017), removing the two paralogous genes was enough to shift the topology inferred with the three codon positions (T1) to the highly supported amino acid topology (T2) in all cases (see Figure 4C). Automatic filtering with TreeShrink was inefficient as it resulted in supporting the artifactual T1 topology (BP = 73) and with an even higher gene concordance factor (CFgene = 28.3 vs. 25.8) than the original nucleotide dataset with all three codon positions included when analysed with a single concatenated model. In contrast, the BLC procedure retrieved the amino acid topology (T2) with strong support (BP = 100), even with a single concatenated model. As shown in Chiari et al. (2012), removing the saturated third codon positions worked for all filtering methods, as was also the case when using a gene partitioned model, which in all cases supported the T2 topology (see Figure 4C). However, in the latter case, the NT_123 and NT123_TS datasets only moderately supported the amino acid T2 topology (BP = 78 and 67, respectively), whereas all other methods and datasets provided strong support (BP = 100).

Finally, it is worth noting that the BLC filtering method generally resulted in higher gene- and site-concordance values compared to other filtering approaches (see Figure 4C), thus demonstrating that gene tree incongruence was reduced by efficient sequence filtering in individual gene alignments. Compared to Bayes factors and likelihood calculations, this method provides an automated and computationally efficient approach to decrease the impact of data error on phylogenomic inference. More generally, "gene incongruence" as detected by current methods does not seem to stem from biological processes. Instead, they

were mostly caused by orthology error and saturated positions, two typical methodological issues during phylogenomic data construction and analysis. Our experiment also highlights the importance of the signal-to-noise ratio. When low quality third codon positions were included, more noise than signal was incorporated because it is hard to extract signal from saturated data with the evolutionary models used here (note that codon models could behave differently as they explicitly account for genetic code structure). When only the first and second codon positions were retained, the absolute signal quantity was lower but the signal-to-noise ratio was higher as we discarded the noisy saturated data. Discarding data that would not be correctly modelled is one way of improving the accuracy of phylogenomic analyses.

4.3 Random contamination and short internal branches

Although it has been shown that cross-contamination (Laurin-Lemay et al., 2012; Simion et al., 2018) and contamination (Philippe et al., 2011b) can drastically distort phylogenomic trees, the level of contamination necessary to induce reconstruction errors as well as the nature of the errors are unknown. In this case study, we introduced various amounts of contaminants into a clean dataset and inferred phylogenomic trees to assess the impact of sequence contamination on phylogenomics. We used the eukaryotic reference alignments maintained in the Philippe lab from which we selected 33 molluscan species as well as 15 lophotrochozoan species as a close outgroup. For the contaminant sequences, we selected species from taxonomic groups observed to be frequent contaminant sources in real transcriptomic datasets (Philippe, unpublished observations): Amoebozoa, Apicomplexa, Arthropoda, Choanoflagellata, Ciliophora, Cryptophyta, Deuterostomia, Diplomonadida, Dinophycae, Fungi, Haptophyta, Heterolobosea, Kinetoplastida, Microsporidia, Nematoda, Platyhelminthes, Rotifera, Stramenopiles and Viridiplantae. Some chimaeras between closely related species were made to reduce the amount of missing data (for details, see Supplementary Material website). A total of 110 species were finally selected and the dataset was constructed by Philippe et al. (2011b). Briefly, ambiguously aligned positions were removed using Gblocks (Castresana, 2000) and a supermatrix was assembled using SCaFoS (Roure et al., 2007). Only 143 genes with less than 27 missing species were considered (see Supplementary Material website), yielding an alignment of 30,517 positions from 110 species with 13% missing data. From this alignment, we extracted an alignment of 48 uncontaminated data (the 33 molluscan and 15 lophotrochozoan species) with 25% missing data, which was used as reference.

For the sake of simplicity, the protocol is described for a contamination level of 5% of species and 5% of genes. The same protocol was repeated for all combinations between [5, 10, 25, 50] percent of genes and [5, 10, 25, 50] percent of species. Ten replicates were done for each contamination level. Briefly, 5% of genes (i.e. 7 genes) were randomly selected over the 143 proteins in the dataset. For each selected gene, 1 to 5% of the 48 species (i.e. 1 to 2 species) were randomly selected as contamination targets and we randomly drew a value between 1 and 5% to mimic the fact that in real alignments the contamination level varies greatly among species. Note that the target species were different for each gene. For each sequence to be contaminated, a species was randomly selected among the remaining 62 non lophotrochozoan species, and its sequence was used to replace the original target sequence. A supermatrix was then assembled using SCaFoS (Roure et al., 2007), yielding an alignment of the same size and level of completeness as the uncontaminated alignment (30,517 positions, 25% missing data).

The accuracy of the phylogenetic inferences performed in the presence of contamination

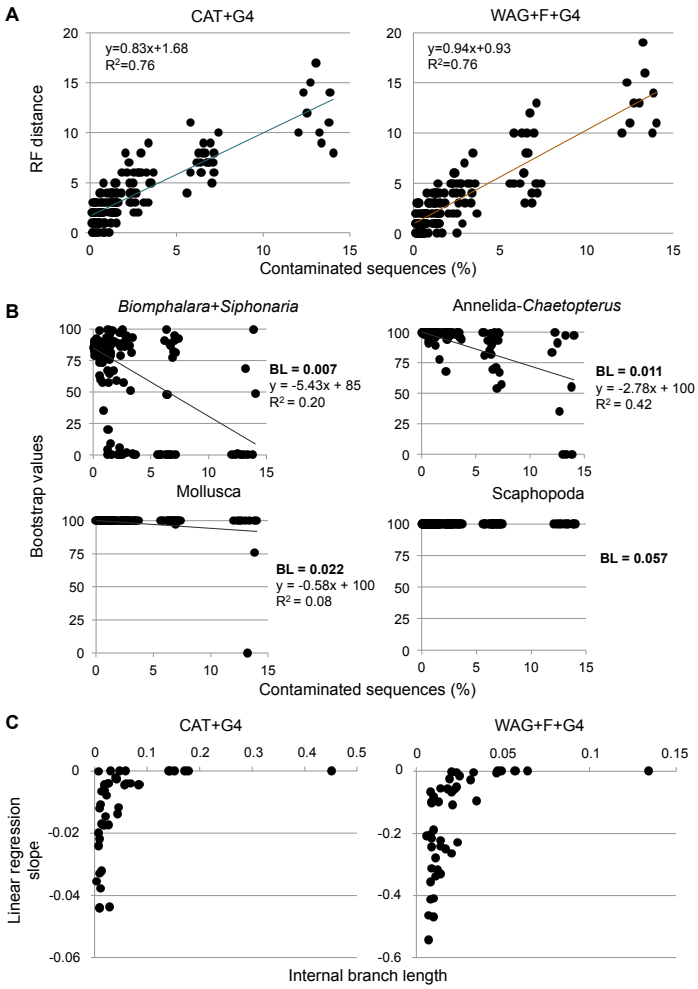


Figure 5 Effect of an increasing contamination level on the phylogenetic accuracy. A) The Robinson-Foulds distance was measured against the topology obtained from the uncontaminated dataset for all contaminated datasets. The inferences were conducted with the CAT+G4 and WAG+F+G4 models. The linear regression is plotted for each diagram. B) For four branches of different length, the bootstrap values of the inferences performed with the WAG+F+G4 model are plotted against the percentage of contaminated sequences. The linear regression is plotted for each node. BL is the internal branch length as estimated by the WAG+F+G4 model. C) The slope of the linear regression between the statistical support (as dependent variable) and the contamination level (as explanatory variable) is plotted against the internal branch length for all branches present in the topology obtained from the uncontaminated dataset with the CAT+G4 and WAG+F+G4 models.

was measured using the Robinson-Foulds distance (Robinson and Foulds, 1981) against the tree inferred from the dataset without contamination. As expected, the accuracy decreased with the amount of contaminated sequences, but no difference was detectable with respect to the sequence evolution model, with similar performances obtained with the CAT+G4 and WAG+F+G4 models (see Figure 5A).

As shown in Figure 5B in the case of the WAG+F+G4 model, the statistical support for a given node decreased with the contamination level. As expected, contamination had a greater impact for short branch lengths, while being negligible for medium and long

branches (compare slope and BL values on Figure 5B). Similar results were obtained with the CAT+G4 model (data not shown). To further validate this result, the linear regression slope (as in Figure 5B) was plotted against the branch length for all non-trivial bipartitions present in the trees inferred without contamination (see Figure 5C). For branches with a length greater than 0.1 (CAT) or 0.04 (WAG), the node was almost always recovered with maximal support and the slope was close to 0 (see Figure 5C). Otherwise the slope decreased with the branch shortness, indicating that inference was on average more sensitive to contamination with shorter branches. The dispersion of the slope values for a given branch length was likely due to the other parameters influencing phylogenetic inference (e.g. depth in the tree, number of taxa in the bipartitions, heterogeneity of evolutionary rates of species surrounding the branch). In summary, at realistic random contamination levels (<5%), only short branches will be negatively affected while most of the topology will remain unchanged. Note that this was due to the randomness of this experiment, thus producing noise. Had we simulated a contamination pattern consistently affecting the same taxa, thus introducing non-phylogenetic signal, even medium and long branches would have likely been affected. In conclusion, the main effect of a limited level of random contamination is a small reduction in the phylogenomics statistical power, so only frequent and biased contamination could explain incongruent phylogenomic trees.

4.4 Reappraisal of phylogenomic signal dissection methods

Recent years have brought forth a particular kind of phylogenomic analysis that aims at using single genes or single alignment sites to investigate phylogenetic relationships that are notoriously hard to resolve. We will here refer to these as "Constrained Topology Analyses" (CTA) as they include slightly different analyses schemes (e.g. Gene Genealogy Interrogation (GGI) [Arcila et al. 2017], ΔGLS and ΔSLS [Shen et al. 2017], Maximum Gene-Wise Edge (MGWE) [Walker et al. 2018], Bayes Factors [Brown and Thomson 2017]). These approaches use the general idea underlying all supertree methods that a "majority vote" from many small data subparts will help determine the best tree. They measure congruence and conflict of genes (or sites) relative to constrained tree topologies and then relies on a "majority vote" to determine which of these topologies is best supported by the data (see also Chapter 3.4 [Bryant and Hahn 2020]). CTA approaches were recently used to assess a variety of phylogenetic relationships, including the position of ctenophores (Arcila et al., 2017) (Shen et al., 2017), otophysans (Arcila et al., 2017), turtles (Brown and Thomson, 2017) (Walker et al., 2018), carnivorous Caryophyllales (Walker et al., 2018), or *Amborella* within angiosperms (Smith et al., 2015).

In this section, we reappraise the potential of CTA approaches to resolve notoriously difficult phylogenetic relationships while criticizing the scant amount of phylogenetic signal they rely on and highlighting the ever-important problem of model fit. CTA approaches are based on the assumption that every single-gene analysis yields enough phylogenetic signal to inform the difficult node under scrutiny. Since these nodes usually correspond to ancient events and/or short internal branches, it can be doubted that the findings of a single-gene analysis could reliably support a given topology versus alternative ones. Indeed, all recent CTA studies have reported a very high number of uninformative genes, and 82.8% to 97.3% of the genes did not reliably support either of the two main topologies under scrutiny (see Figure 6A). The data subparts are generally too small to resolve the problem at hand. This, in itself, is not problematic and is reminiscent of the era of single-gene phylogenetics where all phylogenetic analyses were interpreted in light of these limitations. In CTA, however, all of these single-gene analyses are often genuinely combined and the majority solution is con-

A

	Genes not preferring any of 2 main topologies	Total Genes	Genes supporting the best topology	Uninformative Genes	Informativeness Criteria
Smith et al. 2015	94,3%	5863	204	4024	Bootstrap > 50
Walker et al. 2017	82,8%	93	7	55	SH-like > 80
Arcila et al. 2017 (Metazoa)	96,2%	209	7	201	AU test p-value < 0,05
Arcila et al. 2017 (Eutheria)	97,3%	414	7	403	AU test p-value < 0,05
Arcila et al. 2017 (Neoavia)	93,8%	259	9	243	AU test p-value < 0,05
Shen et al. 2017	na	209	na	na	na

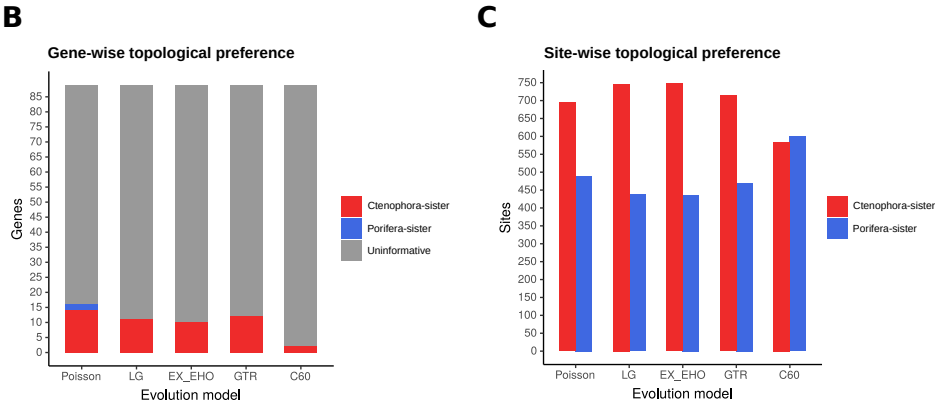


Figure 6 The limitations of CTA approaches. A) Summary statistics of recent CTA studies focused on hard to resolve phylogenetic nodes. Values in the analysis conducted by Smith et al. (2015) are means of the corresponding values for the three deep nodes under study (nodes 3, 4 and 6). B and C) Comparative reanalyses of a previous study (Shen et al., 2017) at gene and amino-acid site levels. B) Gene-wise information detected by CTA under several evolution models. Uninformative genes did not significantly reject the other topology tested according to the AU topology test (Shimodaira, 2002). C) Site-wise information detected by CTA under several evolution models when using the 5% strongest sites (as in the original study of Shen et al. 2017).

sidered to be the best. This is problematic for at least two reasons: (i) there is no guarantee that the majority signal is phylogenetic, as it could stem from an error source that would be additive in nature and therefore lead to a substantial systematic error (i.e. non-phylogenetic signal), and (ii) single gene inference is more complicated because less information is available to accurately infer the evolutionary model parameters. For example, if LBA artefacts affect many single-gene analyses, then the CTA results will erroneously consider that the majority LBA species tree is the correct topology. These issues are even more problematic when CTA is used at the scale of single site (see Shen et al. 2017). Using per-site likelihood differences between contradicting topologies is very similar to counting synapomorphies to build the best species tree: it would only work in the presence of low or negligible homoplasy, which is clearly not the case in large phylogenomic datasets. Historically, questions regarding dataset size and phylogenetic signal extraction triggered the emergence of "total-evidence", "multi-gene" and later phylogenomic approaches (Delsuc et al., 2005). This led to the resolution of many parts of the tree of life by maximising the amount of phylogenetic signal that could be extracted from the data. In this regard, recent CTA methods can be viewed as a return to a "low-evidence" approach as well as a violation of the ideal joint model described at the beginning of this chapter (see Figure 1).

CTA approaches can thus theoretically support an erroneous species tree, like conventional phylogenomic approaches. Since they are based on the sum of single-gene or single-site

signals, they are also impacted by the practical difficulties inherent to meeting phylogenetic objectives: using high quality data under a satisfactory evolutionary model in order to infer an accurate gene genealogy. The importance of data quality was stressed in Sections 4.1 and 4.2. We thus tested the impact of model choice on CTA approaches at gene and site scales by reanalysing a dataset from a recent study focused on the position of ctenophores within metazoans (i.e. the "D16_Opisthokonta" dataset of 89 genes, see Shen et al. 2017). This latter study originally did not check the significance of gene support for a given topology, so we used the same approach as another study instead, based on the approximately unbiased topology test (Arcila et al., 2017). Genes that do not significantly support any given topology are hereby called "uninformative". In addition, we checked the potential test rejection of *a priori* constrained topologies compared to the unconstrained, genuine, gene topology (i.e. *star tree*). This test has not been conducted in recent CTA studies, and our results shows that a large proportion of constrained topologies are significantly rejected when compared to the unconstrained topology, across evolution models: 98% (Poisson), 97% (LG), 98% (EX_EHO), 93% (GTR), and 93% (C60), see Supplementary Material website. This confirms that most single-gene datasets do not carry enough phylogenetic signal to recover a reasonable species tree, regardless of the evolution model used.

Our gene-wise reappraisal of CTA approaches using different sequence evolution models revealed the same trend as noted in previous studies: a very large amount of the genes are uninformative as they do not contain enough phylogenetic signal to favour any constrained topology (grey bars in Figure 6B). Moreover, improving the model complexity decreases the number of informative genes, suggesting that these genes might support a given topology under a simpler model because of model misspecifications. With the C60 site-heterogeneous model, virtually all genes are considered as uninformative regarding the two topologies under scrutiny (Figure 6B). The impact of the model choice is even greater when considering site-wise data. The absolute site-preference for the two topologies tested decreases with the model complexity, as for gene-wise analyses (see AU tests results provided in Supplementary Material website), again indicating that general site support for topologies might be overestimated with simple models. More importantly, whereas simple models show a greater number of sites seemingly supporting ctenophores as a sister-clade to other metazoans (as in Shen et al. 2017), using the more complex and better fitting C60 site-heterogeneous model leads to the opposite conclusion and favours sponges as a sister-clade to other metazoans (see Figure 6C). This highlights that CTA approaches can be highly impacted by model choice upon which the conclusions of studies are based on.

The critical reappraisal of CTA approaches presented here is not intended to discourage its potential use for phylogenomics. Dissecting conflicting signals in large heterogeneous phylogenomic datasets is both important and helpful. However, our results show that CTA still needs to be refined in the following areas: i) investigating the validity of using a constrained topology when the data significantly supports a different topology, ii) ensuring data quality, and iii) selecting the most adequate evolution model for each data subset. It is still doubtful that CTA could effectively resolve notoriously contentious relationships because it is hard for the models to extract enough phylogenetic signal from a small dataset (e.g. see Wang et al. 2019). Small datasets indeed contain a limited absolute amount of phylogenetic signal and the lack of data hampers complex models from accurately estimating parameter values. The main results of CTA studies to date are that outlier genes that generally channel phylogenomic analyses towards an erroneous species tree actually (Brown and Thomson, 2017) or likely (Walker et al., 2018) correspond to data errors (e.g. paralogs, contamination). In phylogenomics studies, CTA approaches might therefore be an efficient data quality

check tool for detecting outliers, rather than being an efficient phylogenetic signal extraction approach.

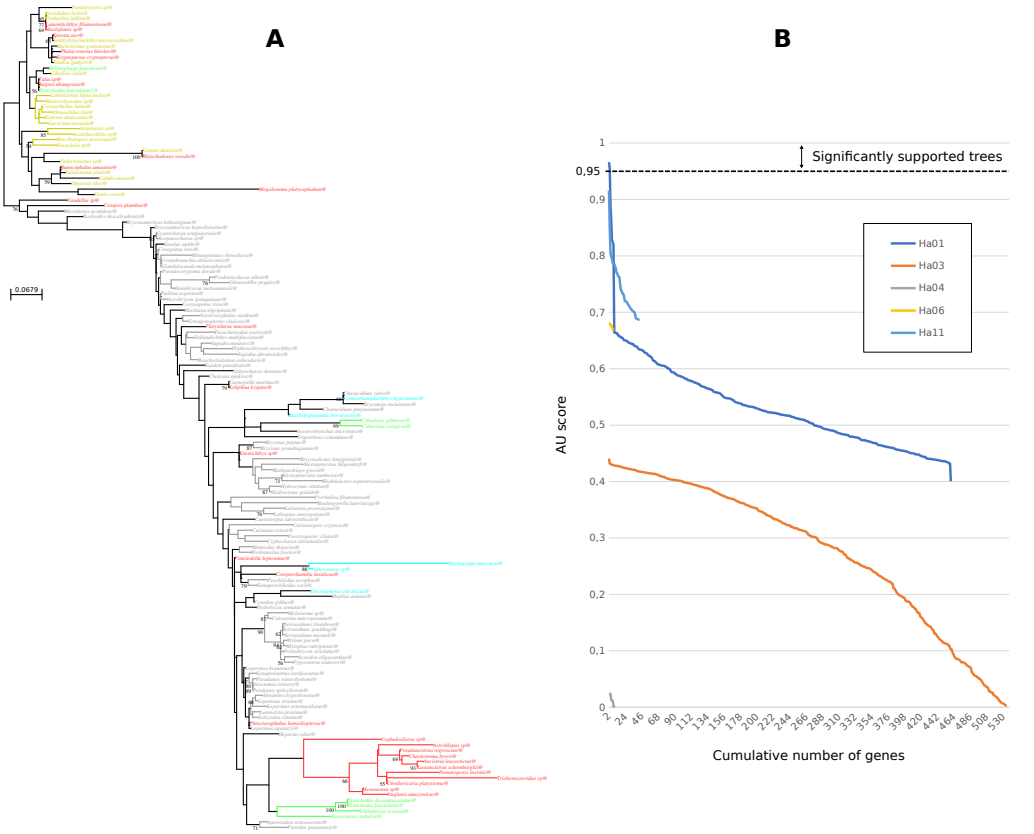


Figure 7 Example of cross-contaminations in the dataset of Arcila et al. (2017). The alignment of the ENSDARG00000061941_5_17739391_17739087 locus was analysed under the GTR+G model using RAxML. A) To illustrate cross-contamination simply, the five major monophyletic groups are shown in color (Cypriniformes in yellow, Gymnotiformes in blue, Siluriformes in red, Characoidei in grey and Citharinoidei in green). For instance, the sequence of the Characoidei, *Characidium zebra*, is identical to *Gymnorhamphichthys hypostomus*, a member of Gymnotiformes. It is likely that the correct sequence is that of *Characidium zebra* since it is closely related to *Characidium purpuratum*. B) Reanalysis of Arcila et al. (2017) dataset using IQ-TREE to perform the Approximately Unbiased (AU) topology test. Lines represent the cumulative number of genes (x axes) supporting each hypothesis with highest probability (rank 1) and their associated P-values (y axes) according to the AU topology test. Values above the dashed line indicate all rank 1 hypotheses that are significantly better than the alternatives ($P < 0.05$), whereas those below the dashed line are also rank 1 but without statistical significance.

4.5 Opening the phylogenomic black box

Phylogenomic pipelines and their various components are becoming increasingly complicated, and the volume of data to handle is exponentially increasing. Accordingly, it is tempting to consider the pipeline as a black box and to focus on the last step, i.e. alignment-based tree inference. However, the road from biological samples to the species tree is long and paved with multiple potential errors, as shown in this chapter. Placing too much trust in the phylogenomic black box is thus risky. For instance, Smith et al. (2011) stated that "Phy-

loBayes misidentified the data type of [their] matrix", without questioning their pipeline or looking at the data, despite the fact that six amino acids (E, F, I, L, P and Q) were entirely missing from their dataset. This led the authors to publish a corrigendum to their study stating that this error was due to a bug in their phylogenomic pipeline (Smith et al., 2013). Similarly, Finet et al. (2010) inferred that two well-established clades (Zygnematales and Coleochetales) were not monophyletic, although their study focused on a sister-group of land plants. Yet a close visual examination of their alignments and corresponding gene trees revealed numerous cross-contamination events that were responsible for these unexpected results (Laurin-Lemay et al., 2012).

It is thus time to open the phylogenomic black box, take a close look at the intermediate data and results, and check that they make sense in the light of current knowledge. We briefly illustrate this by dissecting a recent study on the radiation of Otophysi fish (Arcila et al., 2017). The relationships between four well-established clades (Gymnotiformes, Siluriformes, Characoidei and Citharinoidea) were recognized as difficult to resolve. The authors sequenced 1,051 genes for 225 species, using monophyletic Cypriniformes as outgroup. They used 45 different tree reconstruction methods based on concatenation and species tree approaches and observed that only 5 out of the 15 possible topologies were never supported. The two most frequently recovered topologies were retrieved in only 11 and 9 of the 45 results. In contrast, the monophyly of the 5 clades (Cypriniformes, Gymnotiformes, Siluriformes, Characoidei and Citharinoidea) was always recovered with maximal support. This result is in full agreement with previous knowledge: the length of internal branches is long for the monophyly of the 5 clades (plenty of phylogenetic signal) and short with regard to the relationships among the 5 clades (sparse phylogenetic signal). Instead of concluding that the data quantity was insufficient to resolve this radiation, the authors used the GGI approach in an attempt to find signal in single gene trees. The theoretical expectation is that single genes should have some signal in support of the monophyly of the 5 clades and virtually no signal for their inter-relationships. Indeed, if a single gene provides strong support for any inter-clade relationship, this gene likely did not follow the same historical path as the 1,050 other genes (e.g. contains paralogs). Surprisingly, Arcila et al. (2017) disregarded this theoretical expectation and instead decided "to gain additional insights [...] [to] constrain gene-tree space to a small number of relevant options (15 in this case; Figure1) [to] overcome gene tree estimation error." In other words, they constrained branches for which a single gene was expected to have some signal so as to find support for branches for which the single gene was not expected to bear any signal. Quite surprisingly, their approach seemed to succeed since they found 325 genes that strongly supported one topology (topology H_0 which was supported by seven of the 45 analyses described above) and 69 another one (topology H_a10 which was never supported by any of the 45 analyses). This raises two questions: (1) why was monophyly of the 5 clades constrained, and (2) why was there so much phylogenetic signal in 40% of the loci for the short internal branches connecting these 5 clades but none when concatenated genes were analysed? Indeed, it is striking that only one out of 23 concatenation approaches recovered a topology favoured by the gene-scale approach (i.e., H_0 and not H_a10); Instead, concatenations mostly supported H_a01, H_a03 and H_a04. Unfortunately, the answers are to be found in errors in the phylogenomic pipeline, not in any previously overlooked biological properties.

First, Arcila et al. (2017) recognised that: "it is possible that deep coalescences could result in particular gene histories that display non-monophyly of one or more subclades. This possibility, however, is highly unlikely because internal branches subtending subclades span 20–65 million years of evolution." But they did not foresee that the non-monophyly

of the 5 clades could be due to data error (i.e. cross-contamination events). A comparison of single gene and concatenation branch lengths using the BLC method (see previous section 4.2) highlighted numerous anomalous terminal branch lengths and led to some very poor correlation coefficients. For instance, the phylogeny based on the ENSDARG00000061941_5_17739391_17739087 locus revealed numerous cross-contamination events (see Figure 7A). Among other events, the Siluriformes *Tatia* and *Bagrus ubangensis* are identical to the Citharinoidei *Distichodus fasciolatus*1 (distant from other species of the *Distichodus* genus), and this cluster is deeply nested within the outgroup, i.e. Cypriniformes. Using a variety of approaches based mainly on branch length comparison and BLAST similarity, we estimated that about 12.3% of the sequences of this dataset stemmed from contamination events (20,000 out of the 162,555 sequences. For details, see Supplementary Material website). This indicates that the ENSDARG00000061941_5_17739391_17739087 locus is therefore not an exception. Our findings complement those of a later study from the same group (Betancur-R. et al., 2019), which showed that a dataset from a competing study was cross-contaminated but ironically did not check their own previous data. In summary, constraining the monophyly of the five clades in the Arcila et al. (2017)'s study was mainly necessitated by the presence of numerous incorrectly identified sequences.

Second, the presence of a very strong signal for deep relationships in 394 out of the 1,051 loci reported by Arcila et al. (2017) could solely be explained by software error, i.e. a bug in the RAxML version used, as confirmed by Alexis Stamatakis (personal communication). When we performed the computation with IQ-TREE (Nguyen et al., 2015), our results perfectly fit the theoretical expectations: only 2 out of the 1,051 loci significantly supported any of the 15 possible topologies (see Figure 7B). This observation is in agreement with results of Section 4.4. Moreover, contrary to Figure 2a in Arcila et al. (2017), where the best two topologies were H_0 and H_a10, our results suggested that the best two topologies were H_a01 and H_a03. The topology that was the most frequently recovered in the 45 experiments of Arcila et al. (see H_a01 in their Figure 1) displayed the highest average AU test p-value in our results. This was in full agreement with the theoretical expectation, as the accumulation of a very weak signal (phylogenetic or not) over a large number of genes appeared to be the dominant signal in the concatenated tree.

These examples demonstrated how easily errors can arise along the long road leading from biological samples to species trees and lead to erroneous conclusions (i.e. data error and software error in the case of Arcila et al. 2017). Not only a combination of automatic and manual quality control needs to be performed at each step along phylogenomic pipelines, but also all (intermediate) results should be evaluated in the light of theoretical expectations. For instance, a transcriptome assembly with a reasonable N50 but more than 100,000 contigs, or a very unexpected phylogenetic placement or branch length for a taxon should immediately trigger the opening of the phylogenomic black box. Indeed, such a transcriptome likely contains contaminants and/or fragmented sequences, and a surprisingly long branch taxon could be produced by incorrect underlying data.

5 Conclusions

Multiple types of error occur along the long road from organisms to the species phylogeny, all of which are ultimately due to model violations. They generally decrease the resolving power of phylogenomics (e.g. Figure 5), but occasionally increase it in favour of an erroneous solution (e.g. biased contamination or software bugs). Since very few studies have focused on tracking these errors, we have little idea on the extent of their impact on the species

phylogeny accuracy. We argue that it is essential to identify the most damaging errors (e.g. contamination, annotation errors, orthology errors, alignment errors, single gene tree errors, violation of sequence or gene evolution models) and devote more energy correcting them. Note that the extent of the damaging effect might differ markedly depending on the result of interest. For instance, an annotation error might have a limited impact on topology inference, but a huge impact on branch length and positive selection estimates (Di Franco et al., 2019).

Our current knowledge of biology and evolution could guide us in identifying relevant model violations (and the errors they introduced in the divide and conquer approach of the ideal model in Figure 1. For instance, ignoring ILS would likely only be important for very short internal branches, while over-simplified sequence evolution models could be adequate for minor phylogenetic issues. Posterior predictive checks could enhance studies on these intuitions by quantifying data aspects that are the most poorly explained by the model. For example, recent studies have stressed the importance of rampant discordance between gene trees (Hahn and Nakhleh, 2016), while it could be argued that rampant data error is an equally (or more) serious threat.

Two approaches are conventionally used to eschew these model violations. The first one consists in identifying and removing the most problematic data. For instance, cross-contaminants may be removed from transcriptomic data (Simion et al., 2018), poorly aligned regions (Di Franco et al., 2019) or genes/sites that seriously violate the model assumptions (Roure and Philippe, 2011). This approach is not well founded from a statistical standpoint because the data has to be analysed multiple times while removing data instead of developing an adequate model. However, it is computationally quite efficient and seems reasonnable when founded on solid external knowledge. We believe research in that direction should be pursued. The second approach involves the development of better models. Each sub-model may be improved along the phylogenomic pipeline independently, or sub-models could be combined to allow joint analysis (e.g. alignment and phylogeny). We feel that studies should be focused on sequence evolution models (see Chapter 1.4 [Lartillot 2020a]) and on joint inference of gene trees and species trees (as in Boussau et al. 2013). Yet we stress that software error is an emerging threat to the phylogenomic approach and that increasing the model complexity or implementing clever mathematical tricks to accelerate the computation will increase the software error risk. More resources need to be devoted to the development of high quality software.

Finally, in the current Anthropocene age, the question arises as to whether phylogenomics is a past or future science. Unfortunately, at a time when the environmental footprint of humanity should be drastically reduced [IPCC and IPBES reports], enhancing the accuracy of phylogenomics would require a sharp increase in the computational burden (more data and more complex joint models). As an illustrative example, the computational footprint of xenambulacrarian phylogenomics rose from 7T of CO2 in 2011 (Philippe et al., 2011a) to 260T in 2019 (Philippe et al., 2019), roughly equivalent to 137 round-trip flights between New York and Paris. This raises a legitimate question as to whether pursuing biodiversity science under current scientific practices is reasonable.

References

Alié, A., Hiebert, L. S., Simion, P., Scelzo, M., Prünster, M. M., Lotito, S., Delsuc, F., Douzery, E. J. P., Dantec, C., Lemaire, P., Darras, S., Kawamura, K., Brown, F. D., and Tiozzo, S. (2018). Convergent acquisition of nonembryonic development in styelid ascidians. *Molecular Biology and Evolution*, 35(7):1728–1743.

Allio, R., Scornavacca, C., Nabholz, B., Clamens, A.-L., Sperling, F. A., and Condamine, F. L. (2020). Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Systematic Biology*, 69(1):38–60.

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Vesztrocy, A. W., Dalquen, D. A., Müller, S., Telford, M. J., Glover, N. M., Dylus, D., and Dessimoz, C. (2019). OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Research*, 29(7):1152–1163.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L. J., Ko, K. D., Sabaj, M. H., Lundberg, J., Revell, L. J., and Betancur-R, R. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, 1(2):0020.

Ballenghien, M., Faivre, N., and Galtier, N. (2017). Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biology*, 15(1):25.

Ballesteros, J. A. and Hormiga, G. (2016). A new orthology assessment method for phylogenomic data: Unrooted phylogenetic orthology. *Molecular Biology and Evolution*, 33(8):2117–2134.

Bayzid, M. S., Hunt, T., and Warnow, T. (2014). Disk covering methods improve phylogenomic analyses. *BMC Genomics*, 15(6):S7.

Beaulieu, J. M., O'Meara, B. C., Zaretzki, R., Landerer, C., Chai, J., and Gilchrist, M. A. (2019). Population genetics based phylogenetics under stabilizing selection for an optimal amino acid sequence: A nested modeling approach. *Molecular Biology and Evolution*, 36(4):834–851.

Betancur-R., R., Arcila, D., Vari, R. P., Hughes, L. C., Oliveira, C., Sabaj, M. H., and Ortí, G. (2019). Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: The monophyly of characiform fishes. *Evolution*, 73(2):329–345.

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504.

Blanquart, S. and Lartillot, N. (2006). A bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution*, 23(11):2058–2071.

Blanquart, S. and Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution*, 25(5):842–858.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Plessis, L. d., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4):e1006650.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Boussau, B., Szöllősi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330.

Brown, J. M. and Thomson, R. C. (2017). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology*, 66(4):517–530.

Bryant, D. and Hahn, M. W. (2020). The concatenation question. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.4, pages 3.4:1–3.4:23. No commercial publisher | Authors open access book.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4):540–552.

Chiari, Y., Cahais, V., Galtier, N., and Delsuc, F. (2012). Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (archosauria). *BMC Biology*, 10(1):65.

Cornet, L., Meunier, L., Vlierberghe, M. V., Léonard, R. R., Durieu, B., Lara, Y., Misztak, A., Sirjacobs, D., Javaux, E. J., Philippe, H., Wilmotte, A., and Baurain, D. (2018). Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLOS ONE*, 13(7):e0200323.

Crawford Nicholas G., Faircloth Brant C., McCormack John E., Brumfield Robb T., Winker Kevin, and Glenn Travis C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, 8(5):783–786.

Crotty, S. M., Minh, B. Q., Bean, N. G., Holland, B. R., Tuke, J., Jermiin, L. S., and Haeseler, A. v. (2019). GHOST: Recovering historical signal from heterotachously-evolved sequence alignments. *bioRxiv*, page 174789.

Czech, L., Huerta-Cepas, J., and Stamatakis, A. (2017). A critical review on the use of support values in tree viewers and bioinformatics toolkits. *Molecular Biology and Evolution*, 34(6):1535–1542.

Dalquen, D. A., Anisimova, M., Gonnet, G. H., and Dessimoz, C. (2012). ALF—a simulation framework for genome evolution. *Molecular Biology and Evolution*, 29(4):1115–1123.

Darriba, D., Flouri, T., and Stamatakis, A. (2018). The state of software for evolutionary biology. *Molecular Biology and Evolution*, 35(5):1037–1046.

de Vienne, D. M., Ollier, S., and Aguileta, G. (2012). Phylo-MCOA: A fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Molecular Biology and Evolution*, 29(6):1587–1598.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361.

Di Franco, A., Poujol, R., Baurain, D., and Philippe, H. (2019). Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary Biology*, 19(1):21.

Dondi, R., El-Mabrouk, N., and Lafond, M. (2016). Correction of weighted orthology and paralogy relations - complexity and algorithmic results. In Frith, M. and Storm Pedersen, C. N., editors, *Algorithms in Bioinformatics*, Lecture Notes in Computer Science, pages 121–136. Springer International Publishing.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sørensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q.,

and Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–749.

Dunne, M. P. and Kelly, S. (2017). OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations. *BMC Genomics*, 18(1):390.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94:447–462.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1):157.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584.

Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.

Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G., and Pisani, D. (2017). Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Current Biology*, 27(24):3864–3870.e4.

Finet, C., Timme, R. E., Delwiche, C. F., and Marlétaz, F. (2010). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Current Biology*, 20(24):2217–2222.

Fleissner, R., Metzler, D., and von Haeseler, A. (2005). Simultaneous statistical multiple alignment and phylogeny reconstruction. *Systematic Biology*, 54(4):548–561.

Gatesy, J., Sloan, D. B., Warren, J. M., Baker, R. H., Simmons, M. P., and Springer, M. S. (2019). Partitioned coalescence support reveals biases in species-tree methods and detects gene trees that determine phylogenomic conflicts. *Molecular Phylogenetics and Evolution*, 139:106539.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321.

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17.

Hejnol Andreas, Obst Matthias, Stamatakis Alexandros, Ott Michael, Rouse Greg W., Edgecombe Gregory D., Martinez Pedro, Baguñà Jaume, Bailly Xavier, Jondelius Ulf, Wiens Matthias, Müller Werner E. G., Seaver Elaine, Wheeler Ward C., Martindale Mark Q., Giribet Gonzalo, and Dunn Casey W. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences*, 276(1677):4261–4270.

Herman, J. L., Challis, C. J., Novák, A., Hein, J., and Schmidler, S. C. (2014). Simultaneous bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Molecular Biology and Evolution*, 31(9):2251–2266.

Holmes, I. and Bruno, W. J. (2001). Evolutionary HMMs: a bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–820.

Horiike, T., Minai, R., Miyata, D., Nakamura, Y., and Tateno, Y. (2016). Ortholog-finder: A tool for constructing an ortholog data set. *Genome Biology and Evolution*, 8(2):446–457.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., and Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47:D309–D314.

Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J.-Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M., and Philippe, H. (2017). Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nature ecology & evolution*, 1(9):1370–1378.

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., and Scornavacca, C. (2016). ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058.

Kaduk, M. and Sonnhammer, E. (2017). Improved orthology inference with hieranoid 2. *Bioinformatics*, 33(8):1154–1159.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Koski, L. B. and Golding, G. B. (2001). The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*, 52(6):540–542.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Kozlov, A. M. and Stamatakis, A. (2020). Using raxml-ng in practice. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.3, pages 1.3:1–1.3:25. No commercial publisher | Authors open access book.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24.

Kuzniar, A., [van Ham], R. C., Pongor, S., and Leunissen, J. A. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, 24(11):539 – 551.

Laetsch, D. R. and Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research*, 6:1287.

Lartillot, N. (2020a). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. (2020b). Phylobayes: Bayesian phylogenetics using site-heterogeneous models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.5, pages 1.5:1–1.5:16. No commercial publisher | Authors open access book.

Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–2288.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Laumer, C. E. (2018). Inferring ancient relationships with genomic data: A commentary on current practices. *Integrative and Comparative Biology*, 58(4):623–639.

Laurin-Lemay, S., Brinkmann, H., and Philippe, H. (2012). Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology*, 22(15):R593–R594.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189.

Lowe, C. and Rodrigue, N. (2020). Detecting adaptation from multi-species protein-coding dna sequence alignments alignments. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.5, pages 4.5:1–4.5:18. No commercial publisher | Authors open access book.

Mai, U. and Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, 19:272.

Miele, V., Penel, S., and Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, 12:116.

Minh, B. Q., Hahn, M. W., and Lanfear, R. (2018). New methods to calculate concordance factors for phylogenomic datasets. *bioRxiv*, page 487801.

Mirarab, S., Bayzid, M. S., and Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65(3):366–380.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Monnahan, P. J., Michno, J.-M., O'Connor, C. H., Brohammer, A. B., Springer, N. M., McGaugh, S. E., and Hirsch, C. N. (2019). Using multiple reference genomes to identify and resolve annotation inconsistencies. *bioRxiv*, page 651984.

Moroz, L. L., Kocot, K. M., Citarella, M. R., Dosung, S., Norekian, T. P., Povolotskaya, I. S., Grigorenko, A. P., Dailey, C., Berezikov, E., Buckley, K. M., Ptitsyn, A., Reshetov, D., Mukherjee, K., Moroz, T. P., Bobkova, Y., Yu, F., Kapitonov, V. V., Jurka, J., Bobkov, Y. V., Swore, J. J., Girardo, D. O., Fodor, A., Gusev, F., Sanford, R., Bruders, R., Kittler, E., Mills, C. E., Rast, J. P., Derelle, R., Solovyev, V. V., Kondrashov, F. A., Swalla, B. J., Sweedler, J. V., Rogaev, E. I., Halanych, K. M., and Kohn, A. B. (2014). The ctenophore genome and the evolutionary origins of neural systems. *Nature*, 510(7503):109–114.

Necsulea, A. (2020). Phylogenomics and genome annotation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.1, pages 4.1:1–4.1:26. No commercial publisher | Authors open access book.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment11edited by j. thornton. *Journal of Molecular Biology*, 302(1):205–217.

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., Wallberg, A., Peterson, K. J., and Telford, M. J. (2011a). Acoelomorph flatworms are deuterostomes related to xenoturbella. *Nature*, 470(7333):255–258.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011b). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLOS Biology*, 9(3):e1000602.

Philippe, H., Chenuil, A., and Adoutte, A. (1994). Can the cambrian explosion be inferred through molecular phylogeny? *Development*, 1994:15–25.

Philippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. (2005). Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, 36(1):541–562.

Philippe, H., Poustka, A. J., Chiodin, M., Hoff, K. J., Dessimoz, C., Tomiczek, B., Schiffer, P. H., Müller, S., Domman, D., Horn, M., Kuhl, H., Timmermann, B., Satoh, N., Hikosaka-Katayama, T., Nakano, H., Rowe, M. L., Elphick, M. R., Thomas-Chollier, M., Hankeln, T., Mertes, F., Wallberg, A., Rast, J. P., Copley, R. R., Martinez, P., and Telford, M. J. (2019). Mitigating anticipated effects of systematic errors supports sister-group relationship between xenacoelomorpha and ambulacraria. *Current Biology*, 29(11):1818–1826.e6.

Pich, O., Muiños, F., Sabarinathan, R., Reyes-Salazar, I., Gonzalez-Perez, A., and Lopez-Bigas, N. (2018). Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes. *Cell*, 175(4):1074–1087.e18.

Prous, M., Lee, K. M., and Mutanen, M. (2020). Cross-contamination and strong mitonuclear discordance in empria sawflies (hymenoptera, tenthredinidae) in the light of phylogenomic data. *Molecular Phylogenetics and Evolution*, 143:106670.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Ranwez, V. and Chantret, N. (2020). Strengths and limits of multiple sequence alignment and filtering methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.2, pages 2.2:1–2.2:36. No commercial publisher | Authors open access book.

Ranwez, V. and Delsuc, F. (2020). Accurate alignment of (meta)barcoding datasets using macse. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.3, pages 2.3:1–2.3:31. No commercial publisher | Authors open access book.

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., and Delsuc, F. (2018). MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, 35(10):2582–2584.

Redelings, B. D. and Suchard, M. A. (2005). Joint bayesian estimation of alignment and phylogeny. *Systematic Biology*, 54(3):401–418.

Rey, C., Veber, P., Boussau, B., and Sémon, M. (2019). CAARS: comparative assembly and annotation of RNA-seq data. *Bioinformatics*, 35(13):2199–2207.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107(10):4629–4634.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.

Roure, B. and Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evolutionary Biology*, 11(1):17.

Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evolutionary Biology*, 7(1):S2.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J. M., Guo, Y., Hériché, J.-K., Hu, Y., Kristiansen, K., Li, R., Liu, T., Moses, A., Qin, J., Vang, S., Vilella, A. J., Ureta-Vidal, A., Bolund, L., Wang, J., and Durbin, R. (2008). TreeFam: 2008 update. *Nucleic Acids Research*, 36:D735–D740.

Scornavacca, C., Belkhir, K., Lopez, J., Dernat, R., Delsuc, F., Douzery, E. J. P., and Ranwez, V. (2019). OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Molecular Biology and Evolution*, 36(4):861–862.

Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, 1(5):0126.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3):492–508.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7:539.

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J. C., Manuel, M., Philippe, H., and Telford, M. J. (2018). A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biology*, 16(1):28.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, E., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., and Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, 27(7):958–967.

Simmons, M. P., Sloan, D. B., and Gatesy, J. (2016). The effects of subsampling gene trees on coalescent methods applied to ancient divergences. *Molecular Phylogenetics and Evolution*, 97:76–89.

Simmons, M. P., Sloan, D. B., Springer, M. S., and Gatesy, J. (2019). Gene-wise resampling outperforms site-wise resampling in phylogenetic coalescence analyses. *Molecular Phylogenetics and Evolution*, 131:80–92.

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, 15(1):150.

Smith, S. A., Wilson, N. G., Goetz, F. E., Feehery, C., Andrade, S. C. S., Rouse, G. W., Giribet, G., and Dunn, C. W. (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, 480(7377):364–367.

Smith, S. A., Wilson, N. G., Goetz, F. E., Feehery, C., Andrade, S. C. S., Rouse, G. W., Giribet, G., and Dunn, C. W. (2013). Corrigendum: Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, 493(7434):708–708.

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Springer, M. S. and Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94:1–33.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Steinegger, M. and Salzberg, S. L. (2020). Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Biorxiv preprint https://doi.org/10.1101/2020.01.26.920173.

Struck, T. H. (2014). TreSpEx—detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolutionary Bioinformatics*, 10:EBO.S14239.

Susko, E. and Roger, A. J. (2007). On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution*, 24(9):2139–2150.

Tannier, E., Bazin, A., Davín, A. A., Guéguen, L., Bérard, S., and Chauve, C. (2020). Ancestral genome organization as a diagnosis tool for phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.5, pages 2.5:1–2.5:19. No commercial publisher | Authors open access book.

Tavare, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, pages 57–86.

Walker, J. F., Brown, J. W., and Smith, S. A. (2018). Analyzing contentious relationships and outlier genes in phylogenomics. *Systematic Biology*, 67(5):916–924.

Wang, H.-C., Minh, B. Q., Susko, E., and Roger, A. J. (2018). Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Systematic Biology*, 67(2):216–235.

Wang, H.-C., Susko, E., and Roger, A. J. (2019). The relative importance of modeling site pattern heterogeneity versus partition-wise heterotachy in phylogenomic inference. *Systematic Biology*, 68(6):1003–1019.

Wheeler, W. C., Lucaroni, N., Hong, L., Crowley, L. M., and Varón, A. (2015). POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. *Cladistics*, 31(2):189–196.

Whelan, S., Irisarri, I., and Burki, F. (2018). PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics*, 34(22):3929–3930.

Yang, H., Jaime, M., Polihronakis, M., Kanegawa, K., Markow, T., Kaneshiro, K., and Oliver, B. (2018). Reannotation of eight drosophila genomes. *bioRxiv*, page 350363.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314.

Yu, J. and Thorne, J. L. (2006). Dependence among sites in RNA evolution. *Molecular Biology and Evolution*, 23(8):1525–1537.

Zhukova, A., Gascuel, O., Duchêne, S., Ayres, D. L., Lemey, P., and Baele, G. (2020). Efficiently analysing large viral data sets in computational phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.3, pages 5.3:1–5.3:43. No commercial publisher | Authors open access book.

Chapter 2.2 Strengths and Limits of Multiple Sequence Alignment and Filtering Methods

Vincent Ranwez
AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.
vincent.ranwez@supagro.fr
https://orcid.org/0000-0002-9308-7541

Nathalie Chantret
AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.
nathalie.chantret@inra.fr
https://orcid.org/0000-0002-2512-7644

Abstract

Multiple sequence alignment (MSA) is a prerequisite for most phylogenetic analyses. Aligning sequences to unravel residue homology is a challenging task that has been the focus of much attention in recent decades. Research in this field has been extremely active from both theoretical and practical standpoints. Numerous tools have been developed to align sequences and, more recently, to post-process those alignments and filter out their most dubious parts. Whether or not the inclusion of alignment filtering in a phylogenetic pipeline improves the quality of the inferred phylogenies is still debatable.

The goal of this chapter is not to provide an exhaustive list of all tools available to produce or filter an MSA, but rather to cover the limitations of current alignment methods and their causes, to highlight key differences among MSA filtering methods and provide some practical MSA filtering guidelines.

We consider that filtering methods can be subdivided into two main categories. The first one includes methods that filter MSA by entirely removing some sites or sequences from the MSA. The second category contains MSA filtering methods that mask residues and are able to extract some pieces of information from a site or sequence, while disregarding the remaining information—we called these *picky-filtering* methods. In our benchmark, the filtering methods that perform best are, as expected, in the picky category. When inferring phylogenies, MSA filtering impacts the inferred tree topology but it also seems to significantly improve branch length estimations, especially when a picky-filtering method is used.

How to cite: Vincent Ranwez and Nathalie Chantret (2020). Strengths and Limits of Multiple Sequence Alignment and Filtering Methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 2.2, pp. 2.2:1–2.2:36. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Funding This work was supported by the CIRAD UMR AGAP HPC Data Center of the South Green Bioinformatics platform (`http://www.southgreen.fr/`).

1 Introduction

Multiple sequence alignment (MSA) is used for several kinds of molecular analysis such as identifying "specific determining positions" (SDP) involved in interactions (e.g., protein complexes), post-translational modification sites (phosphorylation, glycosylation, etc.) and, of course, phylogeny inference (Thompson et al., 2011). MSA is a crucial step since phylogenetic inference methods assume that residue homology relationships are correctly reflected by the

input MSA (Chapter 2.1 [Simion et al. 2020]). Parsimony scores, tree likelihoods and tree posterior probabilities are meaningless without this assumption.

Not long ago, each newly sequenced DNA fragment was eye checked by expert biologists and MSA software generated multiple alignments of those highly reliable sequences were diligently curated manually. In the past two decades, sequencing technologies have rapidly progressed, while exomes, transcriptomes and even genomes are now routinely sequenced. This rapid increase in dataset sizes has surely improved the reliability of most of inferred phylogenies as they are now based on hundreds, or even thousands, of loci instead of only a handful. When it comes to working with such large datasets, MSA manual curation is no longer a realistic option and, even when tractable, is a questionable practice as it goes against reproducibility. Many tools have been developed to try to automatically curate alignments by removing part of them, not by correcting them. Overly conservative filtering could thus drastically reduce the available phylogenetic signal along with the sequence alignment errors. There is still an ongoing debate as to whether it is better or not to filter sequence alignments prior to phylogeny inferences since some filtering processes may tend to remove too much of the phylogenetic signal along with phylogenetic noise — so the cure could be worse than the disease. Alignment quality checks are hence sometimes simply ignored, while assuming that errors will somehow be averaged and have little impact on the final biological conclusions, with the vast amount of correct data overwhelming the few incorrect data. However, this argument would not apply to downstream analyses based on the mean deviation. For instance, when searching for branches or loci undergoing positive selection based on non-synonymous vs synonymous substitution rate (dN/dS) analyses (see Chapter 4.5 [Lowe and Rodrigue 2020]), curating the alignment is crucial as alignment errors induce false positives.

MSA methods have been developing since the early 1980s, and it is sill a very active field of research, as illustrated by the number of publications with "multiple sequence alignment" in their title (Figure 1). There was a marked increase in the number of publications at the beginning of the 21^{st} century when sequencing methods became more accessible in the labs, thus further increasing the need for alignment methods. The importance of MSA methods is also reflected by the number of citations of the most successful programs (e.g. $>$ 47.800 citations for ClustalW [Thompson et al. 1994] according to the Web of Science).

Details on aligning two sequences are extensively described in most bioinformatics textbooks, but the steps required to go from pairwise alignments to multiple sequence alignments are often only briefly covered, or not at all. In this chapter we propose to outline those steps, while focusing on the underlying assumptions and shortcuts. We believe that this could be useful for users who may otherwise not understand why even the current best MSA software sometimes generate alignments containing obviously erroneous parts. When using pipelines that automatically chain multiple bioinformatics software to simultaneously analyse thousands of loci, operators may easily lose sight of the imperfections in the underlying methods and the need for caution in interpreting the final results. The next section starts by detailing the strengths and limits of alignment methods. In this respect, Section 2 provides an in-depth explanation of how MSA works in a non-algorithmic way while several figures provide visual schematic representations of the key steps. Then, Section 3 presents the key principles behind MSA filtering methods and pinpoints the inherent limitations of some of them. Finally, based on a large biological dataset, Section 4 provides some insight regarding the impact of MSA filtering methods on the inferred phylogenies.

2 Alignment Methods, Strengths and Limits

Intuitively, an MSA method inserts gap characters '-' inside input sequences to produce a set of longer sequences that are all of the same length, such that residues at the same position in different sequences (aligned residues) share some common properties. More formally, an MSA for a set of n sequences $s_1 \dots s_n$ defined with alphabet $\sum$ is a set of n sequences $S_1 \dots S_n$ which are defined on an enriched alphabet $\sum \cup \{\text{'-'}\}$ such that all S_i have the same length L and, $\forall i$, removing '-' from S_i leads to s_i. The main strong point of MSA methods is that they are the only feasible solution to handle the large datasets that are currently being dealt with. They may be imperfect, they may need to be filtered or manually corrected, but despite their limits they constitute the first step in the alignment process. Manually aligning dozens of sequences of several kb in length from scratch is almost unfeasible and much more tedious than manually curating an (imperfect) alignment generated by MSA software package, which generally return alignments that are very good overall. That said, we think it is worth briefly specifying some of the key steps of MSA methods to gain insight into their current limits and explain why we believe it is crucial to manually or automatically check them before going any further into phylogenetic analysis.

2.1 Multiple Sequence Alignment, Multiple Aims, Multiple Truths

The aims must be clarified before designing software to solve a problem. What is the expected output of the software for a given input? Next it is essential to determine whether the software could actually be designed to meet (in a reasonable amount of time) the specific needs. Regarding multiple sequence alignment, things are not as simple as they may seem at first glance. As pointed out by Morisson in 2006, sequence alignments may be done with different objectives in mind and the ideal alignment for one objective may differ from the ideal one for another application (Morrison, 2006). He listed four distinct objectives for

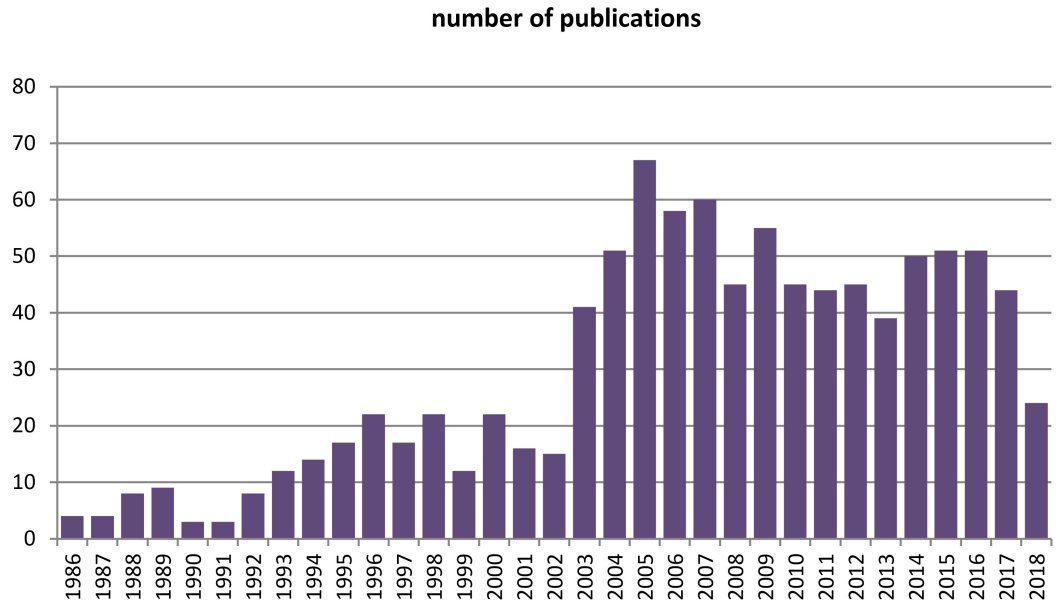


Figure 1 Number of publications including "multiple sequence alignments" in their title. Source: Web of Science.

aligning sequences:

1. Structure prediction that aims to align residues that occupy the same 3D position in the protein
2. Sequence comparison that aims to align preserved functional motifs
3. Database searching that aims to maximize the difference between sequences that are (partially) homologous to the query and those that are not
4. Phylogeny inference that aims to align homologous residues

We have clearly positioned this chapter in the phylogenetic framework for the purposes of the present book. In this context, the aim of an MSA is to match homologous residues together. Thus, residues within the same site (i.e. column) of the MSA are assumed to be homologous, i.e. derived from a common ancestral residue. Given these homology relationships, the plausibility of an evolutionary history (phylogenetic tree) can then be estimated for every site, as well as for the whole alignment (see Chapter 1.1 [Pupko and Mayrose 2020]). For this last step, it is often necessary to assume that sites evolve independently of one another, but this is another story (see Chapter 4.6 [Zou and Zhang 2020]).

2.2 From verbal Aim to imperfect Objective Functions: the big gap

The aim of (phylogenetic related) sequence alignment is thus intimately related to evolution as it is striving to unravel the homology relationships of the residues. That said, the next step towards developing an MSA software is to turn this somewhat abstract objective into a practical measurement of the quality of a candidate alignment with respect to this aim. In computer science, this is called the objective function. The objective function generates a numerical score for each possible solution (here a possible alignment of the input sequences); then we just have to search among all possible solutions to find the one with the highest score. In some fields, the objective function is directly linked to the final objective. For instance, finding a way to place three new antennae in a town to maximize the area receiving a 4G signal; finding the fastest (or shortest) route to go from Paris to Brussels; or finding the longest reading frame in a DNA contig. This is seldom the case in bioinformatics, however, and a large part of the imperfection of bioinformatics method outputs is due to the inability to perfectly transform a biologist's expertise into an objective function with the highest values for the expected output.

An MSA is composed of predicted homologous residues and insertion/deletion events. Hereafter we stipulate how these events are scored (Sections 2.2.1 and 2.2.2) before discussing how pairwise alignments are evaluated based on these scoring schemes (Section 2.2.3), and we then explain how this scoring is extended to MSA (Sections 2.2.4 and 2.2.5).

2.2.1 Cost matrices: principles and limits

As far as amino acid sequences are concerned, the scoring schemes are derived from benchmark alignments that are supposed to be true since they are based on known 3D protein structures and/or have been manually curated. Given those golden standard alignments, the frequency f_{ij} with which two residues r_i and r_j are observed facing each other within a same site (i.e. are homologous) may be compared with the expected frequency $f_i f_j$ of such events, given the two residue frequencies f_i and f_j, if the residue placement is fully random. The log-odds ratio, $log(f_{ij}/f_i f_j)$, is used for this comparison and this is the key principle underlying the construction of a substitution matrix such as the PAM and BLOSUM matrix. The probability of observing two residues at the same site depends on the overall divergence of the considered sequences. There is thus not just one PAM (BLOSUM) matrix but rather a series

of such matrices. Different sequence divergence levels are considered and for each of them a dedicated matrix is built using the subset of the golden standard alignments corresponding to alignments with this degree of sequence divergence. The log-odds ratios are rounded to integer values for the purpose of speeding up the MSA software and reducing the memory space required.

This scoring approach seems mathematically well founded and in agreement with the MSA homology objective, but it has some limits. The learned scores depend on the initial golden benchmark, which could be biased or not representative enough of the sequences to be aligned. Some authors have proposed that substitution matrices could be learned for specific taxonomic groups or sequence types, e.g. for mitochondrial genes (Adachi and Hasegawa, 1996). The LG substitution matrix (Le and Gascuel, 2008) was constructed for phylogeny inference using a broad set of alignments from the (PFAM) protein families database, and the associated web server allows upload of user alignments to learn specific substitution scores for them. Apart from the impact of the initial golden benchmark, the most striking limitation of substitution matrices is that they are scored for a pair of residues and the residue homology is transitive, whereas matrix residue scoring is not. Consider for instance BLOSUM62 scores for the three following residues: phenylalanine (Phe;F), tyrosine (Tyr;Y) and histidine (His;H). Y and F are both aromatic and uncharged and their BLOSUM62 score is +2. Y and H are both polar and hydrophilic and their BLOSUM62 score is +3. However, H and F differ with regard to the four properties and have a BLOSUM62 score of -1. Phylogenetic inference methods using 20 x 20 transition matrices deal with the same problem. The CAT model implemented in phyloBayes (Lartillot and Philippe, 2004) accounts for the diversity of amino acid frequencies among sites, which is poorly captured by classical 20 x 20 matrices. This site heterogeneity is also handled by HMM profiles that are used to model sequence alignment and to search for sequences fitting this model in a database, but as far as we know this feature has yet to be implemented in MSA software.

2.2.2 Gap penalties: principles and arbitrary default choices

Having a protein score matrix is not sufficient to score even a simple pairwise alignment, as the scoring could also penalize gaps inserted to align sequences. The main idea behind gap scoring is that a gap interval (a maximal subsequence of consecutive gap symbols) observed in one sequence results from a deletion of multiple residues in this sequence or from the insertion of as many residues in the other sequences. The penalty (i.e. a negative score that is sometimes called *cost*) for such an event depends on the length of the gap interval, where the penalty increases with the gap interval length. From a biological viewpoint, it seems reasonable that the penalty difference between no gap and a gap of 1 residue would be higher than that between a gap interval of 1 residue and of two residues, which in turn should be higher than the penalty difference between a gap interval of 201 vs. 200 residues. The logarithmic gap cost is in line with this intuitive assumption. According to this gap penalty scheme, the cost for an interval of gaps IG of length $IG[length]$ is $gap_O + log(IG[length])gap_{ext}$, where gap_O is the penalty for the existence of a gap (gap opening penalty) and gap_{ext} is the penalty related to the gap length (gap extension penalty). In practice most MSA methods use an affine gap penalty, where the cost for IG is simply $gap_O + IG[length]gap_{ext}$, where the creation of a new gap interval is penalized more than the extension of an existing one ($|gap_O| > |gap_{ext}|$). This method penalizes the extension of an existing gap interval of one extra residue regardless of the gap interval length. The main reasons for this choice is that, in practice, the affine gap cost leads to faster computation of the alignment cost and that the gain, if any, of using a log affine gap cost is not as clearly established (Cartwright, 2006).

Note that the same gap penalty is used over the entire sequence length whereas it would be reasonable to assume that gaps would be more penalized in some parts of sequences and less penalized elsewhere, e.g. based on the 2D or 3D structure of the corresponding proteins for amino acid sequences (Madhusudhan et al., 2006). There is extensive literature on various gap penalizations and on algorithmic solutions to efficiently compute gaps, but the affine gap cost is by far the most widespread since it is simple, fast to compute and almost as accurate as other more complex penalizations.

Even the simple affine gap cost requires (gap_O and gap_{ext}) parameter setting. Setting the relative cost of those parameters is challenging but adjusting them so that, together with the substitution matrix cost, they will generate a meaningful alignment score is even more challenging. Those two parameters are known to have a strong impact on the output alignments. Without recognised methods to set those parameters, users generally leave them at default values set by developers to perform well on specific MSA benchmarks. Having a method to automatically adjust those parameters to the input sequence dataset is a challenge that has yet to be met although it could drastically improve the MSA accuracy (Wheeler and Kececioglu, 2007). The amount of curated alignments currently available and the progress that has been achieved on machine learning methods enhance the possibilities of learning those parameters via deep learning approaches.

2.2.3 Finding the optimal alignment for two sequences

A two sequence alignment consists of sites where two residues are facing each other (match) and of gap intervals or *indels* (where a sequence of consecutive residues in one sequence is facing gap characters in the other). The overall pairwise alignment score is simply the sum of the match scores (given by the selected cost matrix) plus the gap penalty cost. Figure 2 provides a detailed example of the scoring procedure for an alignment of two sequences S_i and S_j. The overall score of this alignment is +2. If we slightly modify this alignment by moving the amino acid I of S_j in front of the amino acid Q of S_i, we get an alternative alignment that has an overall score of -3 (with the +2 of L facing I being replaced by the -3 of Q facing I). Using this scoring framework, an optimal alignment (one of those with the highest score) can be found efficiently in a time that is proportional to the product of the input sequence lengths. If l_1 and l_2 denote the input sequence lengths, the time needed to solve this problem can be viewed as a polynomial function that tends to be proportional to l_1 times l_2, and the algorithm is said to have a time complexity of $O(l_1 l_2)$. This means that the number of operations needed to solve this problem is of the order of magnitude of $l_1 l_2$. The algorithmic solution, involving dynamic programming, is detailed in many algorithmic/bioinformatics textbook. It is derived from the solution of the problem of identifying the longest common subsequence (LCS) of two sequences, i.e. the longest series of elements that appear in the same order in the two input sequences. This simpler problem, although easy to solve for two sequences, is known to be NP-complete (it cannot be guaranteed that an optimal solution could be found in a reasonable amount of time) for more sequences (Wang and Jiang, 1994). Indeed, an exact solution for n sequences of length l has a time complexity of $O(l^n)$, i.e., requires a number of operations that grows exponentially with respect to the number of sequences, which would be unfeasible for more than a few short sequences.

2.2.4 What is an optimal alignment for more than two sequences?

Defining an objective function that reflects the overall quality of an alignment of several sequences is not straightforward. Let us consider a simple example to understand where

the difficulties lie. Given an alignment of 10 sequences, suppose that a site consists of eight tyrosines (Y) and two histidines (H). If the two sequences containing histidine are sister taxa, then this 8Y2H pattern could be explained by a single Y to H mutation, otherwise two separate mutations are needed. The objective function of an MSA could then be expected to assign a site cost that depends not only on the amino acids present at this site but also on the underlying phylogeny of the corresponding species. The same holds true for gaps, except that things are much more complicated for gaps as they spread along several sites and can overlap so that, even if the underlying phylogeny is known, finding the most parsimonious scenario for gaps would not be that easy.

Given the above remarks, it would be tempting to simultaneously estimate both the MSA alignment and the underlying phylogeny. Many attempts have been made in this direction, i.e. in the parsimony framework (Wheeler, 1996, 2003), as well as in the Bayesian (Herman et al., 2014; Lunter et al., 2005) and likelihood frameworks (Fleissner et al., 2005; Thorne and Kishino, 1992). The POY method (Wheeler, 1996, 2003) was one of the first available methods for simultaneous sequence alignment and phylogeny inference. It was the focus of some attention, as this parsimonious based method seemed to be elegant and perform well. More in depth testing concluded the practical superiority of the conventional two-step approach that first uses MSA software to produce an alignment and then uses it as input for phylogenetic inference (Ogden and Rosenberg, 2007). Indeed, 99.95% of the alignments produced by ClustalW (Thompson et al., 1994) were better than that produced by POY.

In practice, very few of the widely used MSA methods actually account for the underlying evolutionary process, Prank (Loytynoja and Goldman, 2008) being a notable exception.

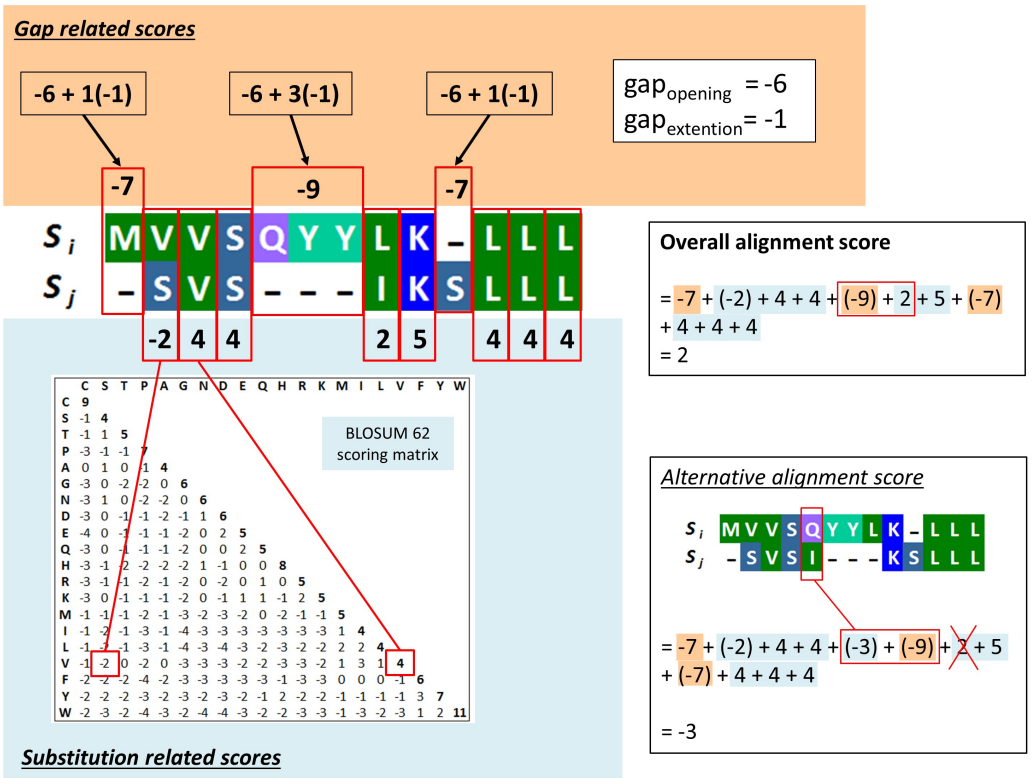


Figure 2 Scoring a pairwise alignment, a detailed example.

There are at least three main reasons for this. Aligning sequences while accounting for the underlying process is harder and much more time consuming overall. Secondly, conditioning the alignment score to an evolutionary scenario can bias the alignment toward this scenario. This latter point is not a problem when the phylogeny is already known or when the main goal of the alignment is not to resolve the phylogeny but rather, for instance, to detect selection footprints. Prank software has proved to be very efficient for this task. In cases where the phylogeny has to be inferred, phylogeny inference routines included within the alignment software are usually much less powerful than dedicated software. For instance, Prank relies on the NJ algorithm, which is a reasonably good distance method but cannot at all compete with up to date probabilistic methods (see Chapters 1.2 and 1.4 [Stamatakis and Kozlov 2020; Lartillot 2020]). There is thus a risk that such methods could generate biased alignments that, by construction, would favour the erroneous phylogeny used to score the alignment (`https://code.google.com/archive/p/prank-msa/wikis/ExplanationDifferences.wiki`). Thirdly, alignments are made separately for each locus whereas phylogenetic inference may consider multiple loci simultaneously.

2.2.5 Objective functions for MSA: the SP-score and its numerous variants

As detailed in the previous section, most MSA software uses an objective function that does not rely on the evolutionary framework. Almost all MSA software uses a variant of the sum of pair scores, or SP-score in short. Given a multiple alignment $\mathcal{A}$, if we consider only its first two rows, then we get a pairwise alignment of the two first sequences aligned in $\mathcal{A}$ that can be scored using, for instance, the pairwise alignment scoring described in Section 2.2.3. The SP-score is basiacally the sum of pairwise scores across all sequence pairs, as illustrated in Figure 3.

More formally, the SP-score of an MSA is obtained by considering all possible pairwise alignments it induces: given two sequences S_i and S_j of the MSA $\mathcal{A}$, the corresponding induced pairwise alignment $(\mathcal{A}|S_i, S_j)$ consists of the two sequences $S_i^{'}$ and $S_j^{'}$ obtained by removing the '-' of S_i (resp. S_j) whenever S_j (resp. S_i) also has a gap at this position/site (see Figure 3 for an example). Conventional algorithms to compute the SP-score of an alignment $\mathcal{A}$, made of L sites and n sequences, proceed by summing up the pairwise scores of its $\binom{n}{2}$ induced pairwise alignments (of length proportional to L) and hence have an overall time complexity of $O(n^2L)$. Faster algorithms are now available for both the general gap penalty and the affine gap cost cases (Ranwez, 2016). Indeed, it turns out that this latter can be solved with a simple and efficient algorithm having a time complexity of $O(nL)$.

An alternative approach is to build an optimal pairwise alignment $\mathcal{A}_{ij}$ for each sequence pair S_i and S_j of the set $\mathbb{S}$ of sequences to be aligned. This is possible as the pairwise alignment problem can be solved in polynomial time. For a given MSA $\mathcal{A}$, the pairwise restriction $\mathcal{A}|S_i, S_j$ can then be compared to the computed optimal alignment $\mathcal{A}_{ij}$, and the closer they are the better it is because the multiple alignment is more consistent with the optimal pairwise alignments. An alternative way of scoring the MSA $\mathcal{A}$, which here we call SP-sym, is thus to sum up the similarity between (a sample of) its $\binom{n}{2}$ induced pairwise alignments and their optimal pairwise counterparts. This scoring scheme was first introduced in T-coffee software (Notredame et al., 2000). Based on this scoring, the searched MSA can be seen as the median of the considered optimal pairwise alignments.

Numerous MSA software packages including Clustal, MUSCLE and MAFFT use the SP-score framework, but they introduce variations in an effort to improve it. Some of them use different substitution matrices for the different pairwise comparisons, while others use a

weighted version of the SP-score, where all pairwise comparisons do not equally contribute to the overall SP-score, the gap scoring can also vary along the sequences based on the 2D structural information or on the gap frequencies in the pairwise alignments of the considered dataset. The most recent releases of MAFFT use a combination of SP-score and SP-sym to build their objective function. Our goal here is not to extensively review those MSA scoring variants but rather to outline the key underlying principles in order to highlight that those scoring schemes are both powerful – each newly launched software package brings new improvements – and imperfect – as they mostly overlook (for sound reasons) the evolutionary relationships of the compared sequences.

2.3 Heuristic search for optimal MSAs: the more you get, the worst the search

Finding an MSA with an optimal SP-score is known to be NP-complete (Wang and Jiang, 1994). An exact solution to this problem can thus only be found when dealing with a small number of short sequences. Hence, all widely used MSA software relies on a heuristic search to find an MSA that has a reasonably high score. This search involves two distinct phases: the first is to build an initial MSA of the input sequences, while the second is to improve this initial MSA through iterative refinement. Both steps extensively rely on the idea of aligning two alignments, i.e., merging two previously aligned, disjoint subsets of sequences into a new, larger alignment.

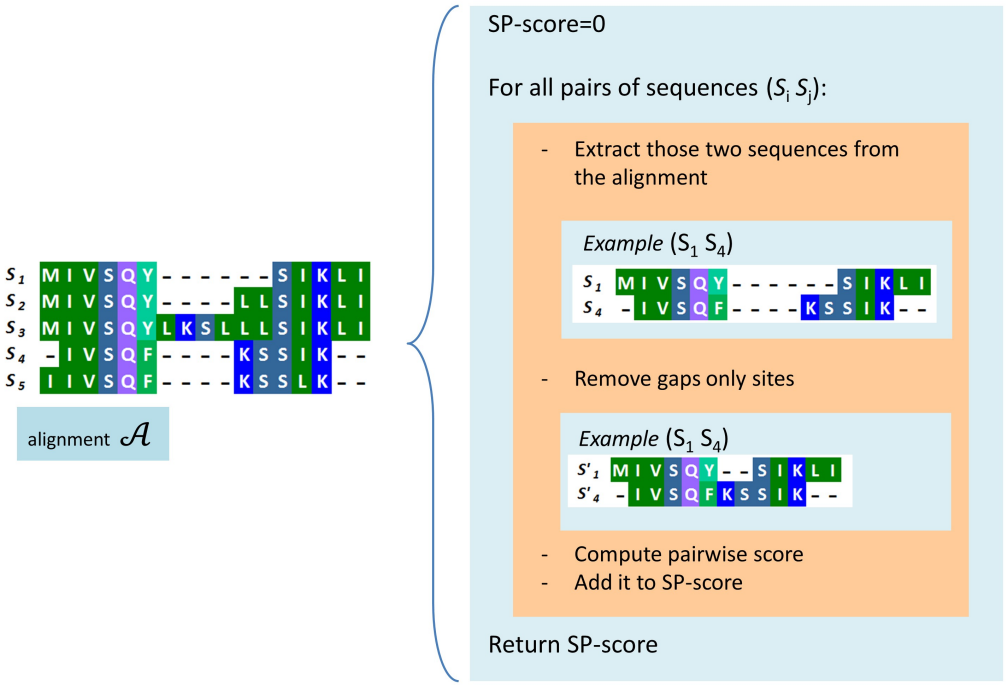


Figure 3 SP-score computation principle for an MSA.

2.3.1 Aligning two alignments and gap penalty approximations

Given an alignment $\mathcal{A}_1$ of a set of sequences $\mathbb{S}_1$ and an alignment $\mathcal{A}_2$ of a disjoint set of sequences $\mathbb{S}_2$, an alignment $\mathcal{A}$ of $\mathbb{S}_1 \cup \mathbb{S}_2$ may be obtained by aligning $\mathcal{A}_1$ and $\mathcal{A}_2$. This alignment task aims to identify pairs of homologous sites of the two input alignments and position them in front of one another in a global alignment. This could be seen as a search for the best alignment of sequences of $\mathbb{S}_1 \cup \mathbb{S}_2$ that respect the homology relationships present in $\mathcal{A}_1$ and $\mathcal{A}_2$, i.e. such that $(\mathcal{A} \mid \mathbb{S}_1) = \mathcal{A}_1$ and $(\mathcal{A} \mid \mathbb{S}_2) = \mathcal{A}_2$.

A naive way to do this, while also showcasing the key idea of the method, is presented below and illustrated in Figure 4. First, for the alignment $\mathcal{A}_1$, consisting of n_1 sequences and l_1 sites, a consensus sequence S_{c1} of length l_1 such that the amino acid at position k of this sequence is one of the most frequent at site k of $\mathcal{A}_1$, is built. Then we proceed similarly to build S_{c2} from the alignment $\mathcal{A}_2$, consisting of n_2 sequences and l_2 sites. In a second step, the consensus sequences S_{c1} and S_{c2} are aligned. In a third step, this resulting alignment is used to derive homologous sites of $\mathcal{A}_1$ and $\mathcal{A}_2$ by assuming that every time two amino acids of S_{c1} and S_{c2} are facing each other the corresponding $\mathcal{A}_1$ and $\mathcal{A}_2$ sites are homologous. To visualize the process, we could imagine that each amino acid of S_{c1} and S_{c2} has a skewer of amino acids attached to it that correspond to the site it summarizes. Furthermore, we imagine that each gap of S_{c1} (resp. S_{c2}) has a skewer of n_1 (resp. n_2) gaps attached to it. Then, given the alignment of S_{c1} and S_{c2}, we can build a global alignment of $\mathcal{A}_1$ and $\mathcal{A}_2$ by merging the two skewers facing each other.

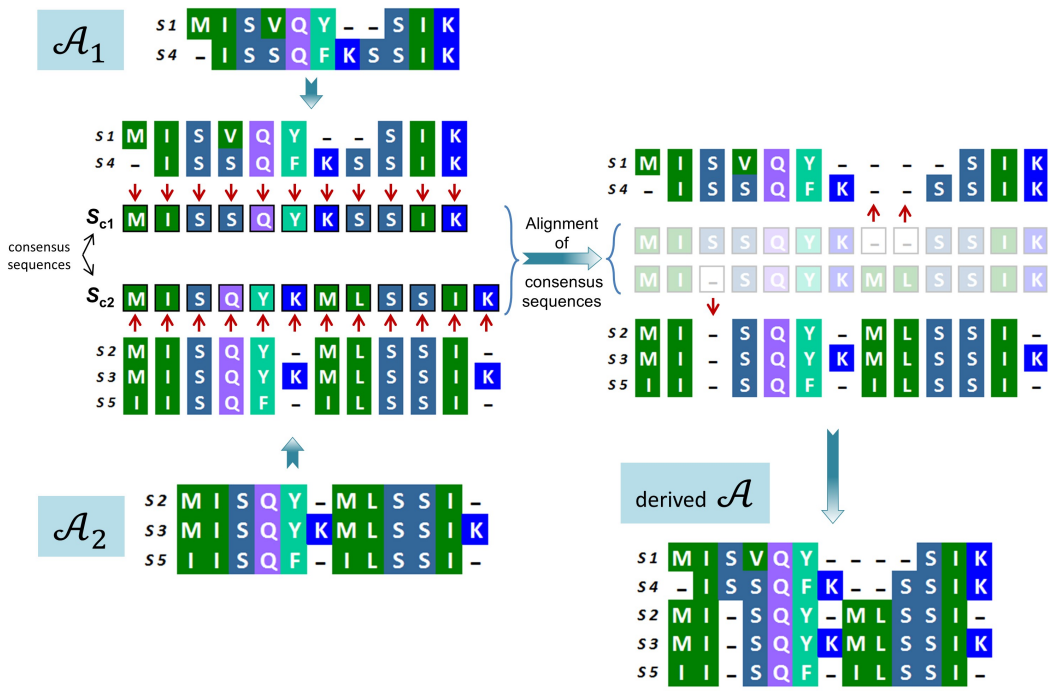


Figure 4 Aligning two alignments, a naive approach.

The naive approach described above is not the one generally used, as summarizing all site information with a single character is a drastic compression that generates a major

information loss. In the example of Figure 4, the third site of $\mathcal{A}_2$ would be better aligned (at least w.r.t. the SP-score) if placed in front of the third one of $\mathcal{A}_1$, but this cannot be done when only the consensus sequences S_{c1} and S_{c2} are used. The dynamic programming approach to achieve optimal two-sequence alignments could be extended to better align two alignments, but it would no longer be certain that the resulting generalization would return an optimal solution. The SP-score of an alignment can be split into three parts: (1) SP_{subst}: the part induced by amino acid homology and computed using the substitution score matrix; (2) SP_{gext}: the part induced by gap extensions; and (3) SP_{go}: the part induced by gap openings. The first two are easy to estimate even when aligning alignments. If we consider putting site i of $\mathcal{A}_1$ in front of site j of $\mathcal{A}_2$, we can readily see how this would impact SP_{subst}, which could be accurately evaluated just by knowing the number of each amino acid present at sites i and j. Similarly, the impact on SP_{gext} may be assessed simply on the basis of the number of gap and non-gap characters at both sites. The situation is much more complicated regarding gap openings since the sites can no longer be considered independently. By considering two consecutive sites, it is possible to determine the upper bound (pessimistic gap count) and lower bound (optimistic gap count) regarding the number of gaps that would be opened by an alignment operation (e.g. placing site i_{k1} in front of j_{k2}), but it is not sure that exact counts of the resulting gap openings would always be obtained (Altschul, 1989).

2.3.2 Building an initial MSA: not every tree is a phylogeny

Several heuristic methods are available to build an initial MSA. The most widespread one involves progressive alignment construction guided by hierarchical sequence clustering (Feng and Doolittle, 1987). This method follows a greedy strategy (the homology established at some point is never questioned afterwards) whereby a larger alignment is built by aligning two smaller ones until all input sequences are jointly aligned. Figure 5 illustrates the procedure, as detailed below.

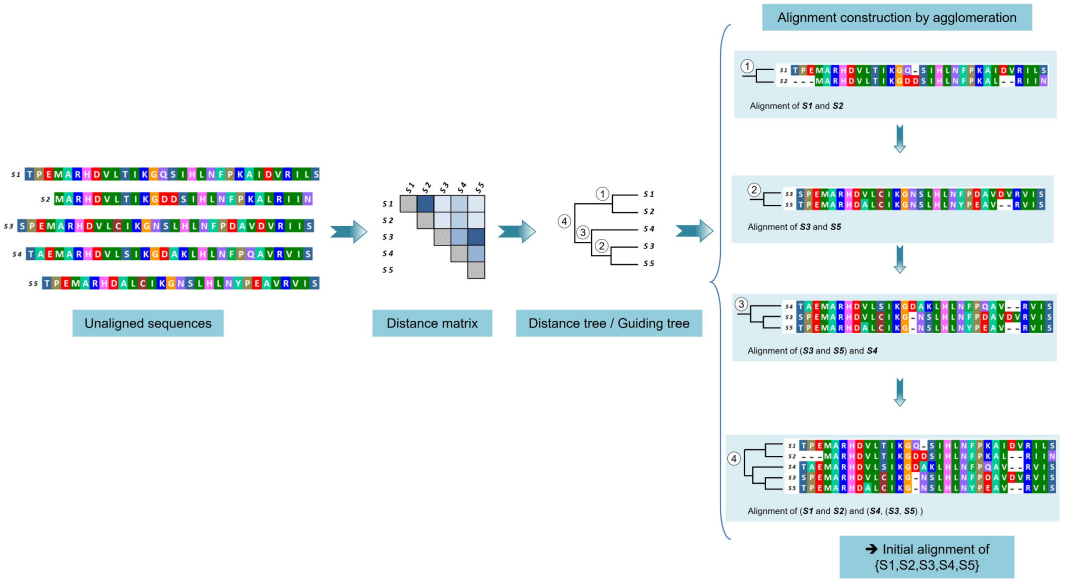


Figure 5 Building an initial MSA using a guiding tree.

The advantage of using this greedy approach is that it is fast, since decisions taken at

one step are never questioned afterwards. The downside is that errors made in the early stages of the process, which condition subsequent choices, may have a devastating impact on the final result. To mitigate this problem, it is thus preferable to start by the easiest tasks that should be less error prone. If two sequences are fully identical, it is really easy to align them and the right solution is much better than any alternative possibility. If the two sequences differ by a single deletion of a few amino acids, the task remains straightforward but there may be some uncertainty concerning the alignment of amino acids at the frontier of the deletion. If one of the sequences is half the length of the other one and no motif longer than four consecutive amino acids is shared by the two sequences, it is quite likely that the alignment will contain (many) errors. The progressive alignment strategy thus first aims to jointly align the most similar sequences as it is an easier and less error prone task. To this end, hierarchical sequence clustering, a so-called guiding tree, is done so as to group the most similar sequences together. Each leaf of this tree is thus associated with an input sequence (i.e. a trivial alignment) and each internal node will then be associated with the alignment of the sequence below it. Those alignments are built by processing the internal tree nodes from tips to the root, which ensures that a node is always processed after its two children. An internal node is simply processed by producing the alignment associated with this node, i.e., by aligning the two alignments of its two children nodes. At the end of the process, the root node is associated with an alignment of the whole sequence set. Note that the guiding tree should not be confused with the input sequence phylogeny. There is surely a link between sequence similarity and species relationships, but most similar sequences are not necessarily derived from the most related species as evolutionary rates can vary during evolution. For a given gene, a human sequence may be more similar to a cow sequence than to a mouse sequence because mouse genes evolve faster. In such cases, it makes sense that the guiding tree pools human and cow sequences so as to postpone the harder task of aligning a divergent mouse sequence with the others, even though humans and mice are known to be closer relatives than humans and cows. In such a situation, the guide tree justifiably differs from the true species tree.

The guiding tree is built based on a distance matrix that provides a measurement of the similarity/divergence between any two sequences of the input set. This similarity can be derived using the pairwise alignment of the two concerned sequences, but performing $\binom{n}{2}$ pairwise alignments is extremely time consuming and most MSA software packages rely on a pairwise distance estimation based on k-mer contents which are much faster to obtain – a k-mer is a set of consecutive k amino acids in the sequence. The sequence similarity is assumed to increase as the proportion of shared k-mer increases. Finally, note that once a first MSA alignment is obtained in this way, some software uses it to derive an improved distance matrix based on the pairwise sequence alignment induced by this first MSA. Then they build a new guiding tree and infer a new MSA using it. This idea of repeating the progressive alignment construction with an improved distance matrix was introduced, to our knowledge, in the first release of MUSCLE (Edgar, 2004). The initial MSA produced at this stage is highly important since not only is it the starting point for the subsequent alignment refinement stage but also its associated guiding tree will also guide the refinement search.

2.3.3 Optimizing the initial MSA and why this is better done with fewer sequences

The progressive alignment step is explained in great detail in many courses and textbooks whereas, despite its importance in practice, the refinement step is often rapidly described at best. Indeed, an alignment produced just using the progressive strategy can be of very poor

quality due to the greedy approach used to build it.

For most software, this refinement step relies on a "hill climbing strategy" to optimize the objective function: variants of the current alignments are produced and each time a variant better than the current solution is encountered it becomes the new current solution. The optimization process, as illustrated in Figure 6, stops when no better variant is found and the current solution thus becomes a local optimum.

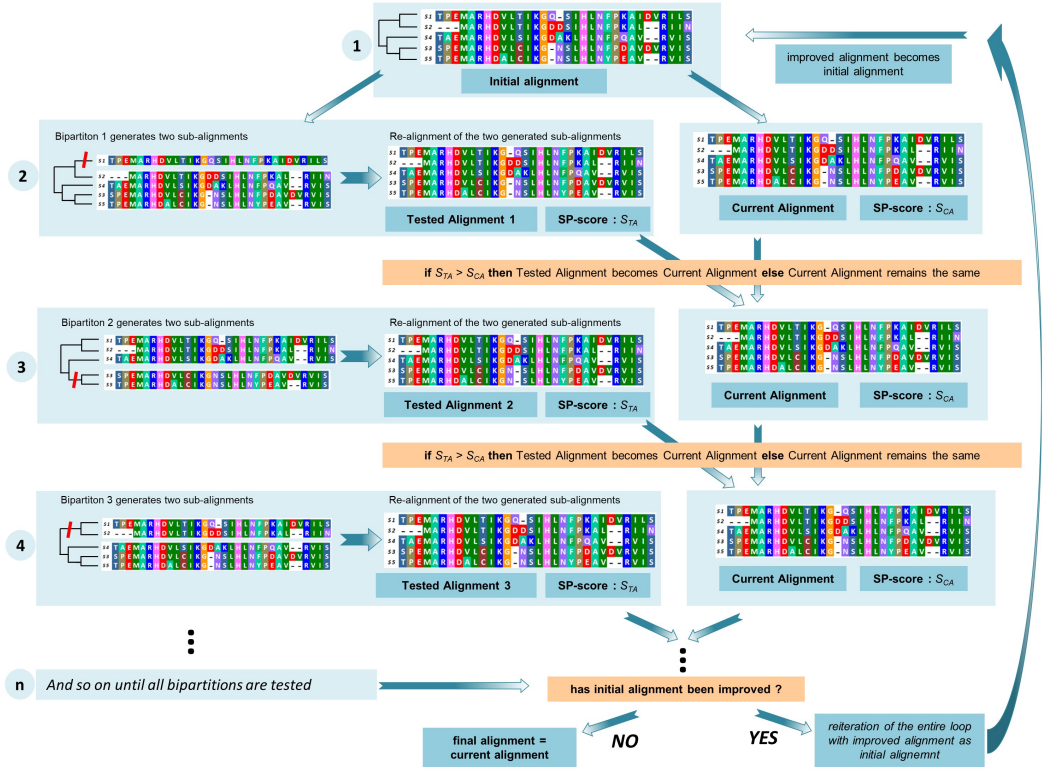


Figure 6 Schematic representation of the 2-cut strategy used to refine an initial MSA.

A maximum number of iterations can also be set to reduce the computation time. The way the variants to test are produced can vary, but in many cases this is done by splitting the alignment in two and realigning the two resulting subalignments. This is sometimes called the 2-cut refinement strategy. Given an alignment $\mathcal{A}$ of a set of sequences $\mathbb{S}$ and a bipartition of $\mathbb{S}$ into $\mathbb{S}_1$ and $\mathbb{S}_2$ ($\mathbb{S}_1 \cup \mathbb{S}_2 = \mathbb{S}$ and $\mathbb{S}_1 \cap \mathbb{S}_2 = \emptyset$), the two induced alignments $\mathcal{A}\,|\mathbb{S}_1$ and $\mathcal{A}\,|\mathbb{S}_2$ are aligned to get a variant of $\mathcal{A}$ that will replace it if and only if it has a better score. As the number of possible bipartitions of $\mathbb{S}$ (2^{n-1}) grows exponentially with the number of sequences n, it is impossible to test all of them, except for datasets with less than a handful of sequences. The guiding tree is then used to define the bipartitions to be tested, and a refinement loop consists of testing all bipartitions corresponding to a branch of the current guiding tree. The current alignment is updated every time an alignment variant with a better score is found, and hence it could be changed several times during each loop. The process stops at the end of a loop if no better alignment has been found, or otherwise a new loop starts. The guiding tree may or not be updated after each loop based on the new best alignment found at that point.

The number of bipartitions/branches within a tree of n sequences is $2n-3$ whereas the

total number of possible bipartitions among n sequences is 2^{n-1}. The fraction of bipartitions that are considered to try to improve the alignment is thus $(2n-3)/2^{n-1}$, i.e. a fraction that exponentially tends toward zero as n grows. For 10 sequences, we test $\sim 3\%$ of all possible bipartitions but only about $\sim 3 \times 10^{-26}$ % for 100 sequences. If some sequences $\mathbb{S}_1$ share a common gap that is misplaced with respect to the other sequences $\mathbb{S}_2$, the chances of precisely testing the alignment variant associated with bipartition $\mathbb{S}_1 \mid \mathbb{S}_2$, and hence correcting the misplaced gaps, are thus almost inexistant for a 100-sequence dataset. This difficulty is illustrated in Figure 7, using a small and simple example where an alignment cannot be improved with the current guiding tree despite the fact that a better alignment obviously exists.

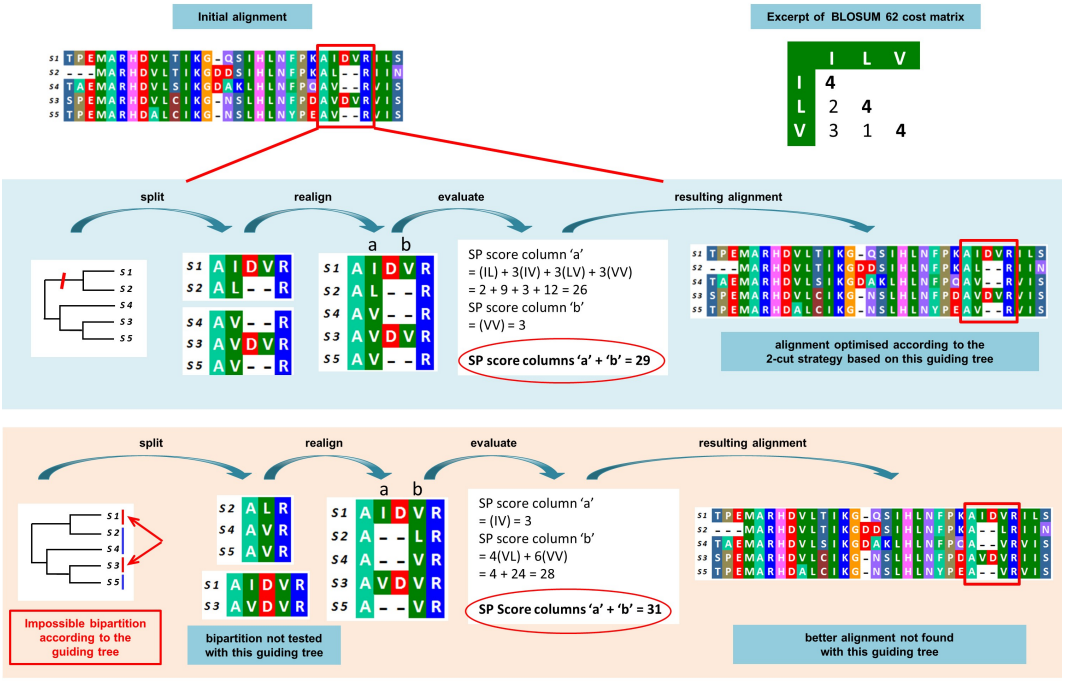


Figure 7 Limits of MSA refinement when bipartitions are made according to the guiding tree. Starting with the initial alignment (top right), the better alignment (depicted on the bottom right) cannot be found with the current guiding tree (depicted on the left) since this tree does not contain the bipartition $\{S1, S3\}|\{S2, S4, S5\}$ that would help find a better alignment.

This somewhat pessimistic view has to be qualified by the fact that, rather than testing randomly picked bipartitions, we test those observed in the guiding tree, which are expected to be more promising bipartitions than random ones. Yet, this phenomenon cannot be overlooked, and is probably one of the reasons that led R. Edgar to state in the MUSCLE 3.8 user guide that "If you have thousands of sequences, then attempting to create a multiple alignment is dubious for many technical reasons. It may be better to cluster first, then align the reduced set of sequences".

3 Filtering alignments, less is more, well more or less

Alignments are the foundation upon which molecular phylogenetic analyses rely. In this part we will consider an alignment obtained by any MSA method as a starting point (alignment methods and optimization were discussed earlier). Once the alignment has been achieved, it intrinsically contains all relevant information that will be used by evolutionary models to reconstruct the history of the aligned sequences, and therefore any error in the alignment could seriously affect the inferred phylogeny.

As explained in the previous sections, MSA methods have numerous shortcomings (they rely on heuristic searches guided by imperfect objective functions). It is thus inevitable that in most cases their output, even if satisfactory overall, is tainted with errors. Many software programs have been designed with the aim of filtering MSAs in order to keep only their most reliable regions. This filtering is done by removing sites, sequences or masking residues (replacing them by the gap symbol '-' or by a symbol representing ambiguity '?', 'N' or 'X'). Filtering MSA would be a reasonable thing to do, but we have to avoid throwing out the baby with the bathwater. It is important to be sure that the filtering does not remove the signal along with the noise caused by the misaligned regions. For MSA filtering, as for any signal filtering process conducted to improve the image or sound quality for instance, the balance between noise reduction and signal loss is key. This balance partly depends on the planned downstream analysis. For example, misaligned regions impacting a single sequence at one time will have little impact on the phylogeny inference, apart from terminal branch length estimations, but they will induce many false positives when searching for loci under positive selection. The efficacy of cleaning methods have been a highly topical issue in recent years with the advent of high-throughput sequencing methods that have dramatically increased dataset sizes, up to the size of whole genomes. MSA manual curation is still applied in small or medium size datasets but is impossible for larger ones. Moreover, the crucial reproducibility issue is another reason for developing automatic MSA cleaning methods.

We consider that filtering methods can be subdivided into two main categories. The first one contains methods that filter MSA by entirely removing some sites or sequences from the MSA. These methods give you only two choices per site and sequence: you either "take it or leave it", which is why we have called them TILI-filtering methods. The second category contains MSA filtering methods that work by masking residues (replacing them by the gap character '-' or by a symbol representing ambiguity '?', 'N' or 'X', depending on the sequence type). Such methods take pieces of information from a site or sequence, while excluding the rest of it, so we have called these picky-filtering methods. Before delving deeper into filtering methods, it would be worthwhile to briefly outline the different cases of poor quality alignment and their causes, along with some possible upstream remedies.

3.1 What are the problematic regions of an alignment?

There are two problematic regions within an MSA, which we call poorly informative regions and wrongly aligned regions. Poorly informative regions are regions where most of the sites contain many gaps. It is hard to say whether the alignment is correct in these regions but, regardless, these regions carry little evolutionary signal. On the other hand, wrongly aligned regions can be defined as regions for which the hypothesis whereby aligned residues would have a common ancestor is unlikely to be correct. Anyone who has ever aligned sequences automatically and carefully looked into the generated alignments, has witnessed such regions. Many alignments include both poorly informative and wrongly aligned regions, or their combination. As specified below, problematic regions may have diverse impacts on the

downstream analysis depending on their characteristics.

3.1.1 Patchy regions

One group of poorly informative alignment regions could be described as 'patchy'. This is when the alignment algorithm has inserted so many gaps that there are long stretches of sites at which gaps predominate (Figure 8a). This situation is obviously caused by the presence of highly divergent regions in the sequence. These patchy regions do not necessarily induce errors in the tree topologies, as they basically do not contain any phylogenetic signals but they may disrupt the bootstrap procedures. Indeed, if too many sites are sampled in those regions for some bootstrap replicates, it may be impossible to compute a tree for those datasets and the inference program may crash without many clues to enable you to spot the problem.

3.1.2 Regions in the vicinity of patchy regions

In the vicinity of patchy regions, you may also find sites for which a greater number of sequences are present (short fragments, like "islands"). In these cases the homology between aligned fragments is often doubtful (Figure 8b). This is a common situation at the 5' and 3' ends of genes, i.e. regions that are known to diverge faster between lineages, in addition to often being subject to erroneous annotations. Such situations are more likely to happen if the number of analyzed sequences is large.

3.1.3 Misaligned regions

In some regions, suboptimal alignments, as shown in (Figure 8c), can be observed (see Section 2.3 for further details). In these cases, any phylogenetic signals will be blurred. These errors are also typically caused by a large sample size in low similarity regions. Indeed, the sequences in the example of Figure 8c were among the most distantly related ones in the dataset, and the algorithm failed to optimize the alignment of those small isolated regions.

3.1.4 Low complexity regions

Repeated characteristics, or small repeated motifs, can lead to another kind of wrongly aligned region, as presented in Figure 8d. The region depicted in this figure is a transmembrane domain, and as such requires a high prevalence of hydrophobic residues. Because of the relatively small number of hydrophobic amino-acids (mainly Ala, Ile, Leu, Met, Phe and Val), this region is particularly prone to reverse substitutions and convergent evolution (homoplasy), resulting in disordered repetitive stretches of hydrophobic amino acids. The resulting nonsense or wrong alignment may distort the phylogenetic signal.

3.2 What causes problematic MSA regions?

Aligning highly similar sequences is quite easy. Note that this idea underlies the greedy MSA strategy which, by using a guiding tree, aligns the most similar sequences first and postpones the harder task of aligning divergent ones. MSA software hence generally produces high quality MSAs when the input sequences are highly similar over their entire length. Difficulties arise when sequences are, locally or globally, highly divergent, possibly due to annotation errors. Short or long motif repeats also cause confusion since it may be hard to

identify homologous motifs, and long indels are particularly challenging for MSA software since an affine gap cost penalty does not fit well such situations.

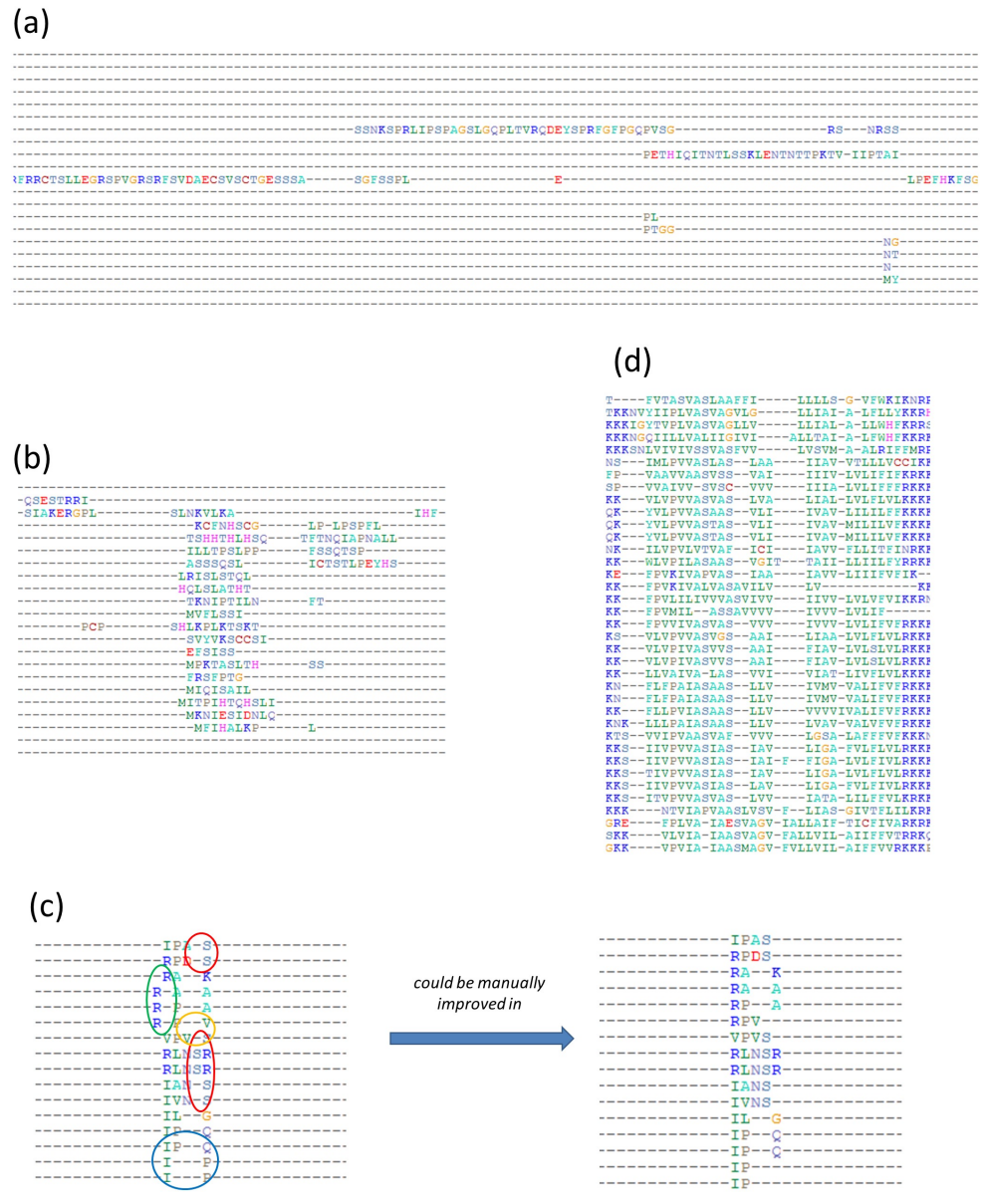


Figure 8 Examples of problematic regions of an MSA.

3.2.1 Highly divergent or non-homologous sequence fragments

As alignment algorithms will always be able to propose an alignment of sequences even when they are only remotely related, the question of the reliability of the produced alignment arises. Note that even when sequences are too divergent, or not even homologous, MSA software will still produce an alignment. However, some phylogeny inference methods will refuse to take this alignment as an input and will produce an error message indicating that the input sequences are too divergent. This occurs, for instance, with most distance methods when the distance matrix, built from the input MSA, contains missing values. Indeed, the pairwise distance cannot be calculated between two sequences that do not share homologous residues according to the input MSA (i.e. no residues placed on the same site). This could indicate that the two sequences are only partially available (the 3' part is missing from one and the 5' end from the other). In this case it is fine to use alternative phylogeny methods (not based on distances). Alternatively, it could also indicate that the two sequences are indeed non-homologous and should not be simultaneously present in this MSA and that this alignment is hence unreliable.

When sequences are highly divergent, they contain little (if any) information about homology relationship between residues. The MSA that could be inferred from such a dataset would hence be almost random, just as would be a phylogeny only inferred based on saturated sites. Even when sequences are generally similar, the question of alignment reliability remains relevant given that in most cases similarity levels are not homogenous all along the sequences being compared. Heterogeneous similarity levels among sequences are often observed when analyzing gene family. Gene families are often defined by the presence of a conserved active domain, while the rest of the protein, since it is a lot less constrained, may evolve at much faster rates and rapidly diverge.

Somewhere along the gradient from highly similar sequences to highly divergent sequences, there is a critical point beyond which to align sequences is not possible, or biologically meaningful - too many substitutions or indels have occurred. Beyond this point, alignment makes no sense since it is impossible to guarantee that the aligned positions are derived from an ancestral state. Just before this point, computational limitations of alignment methodologies induce errors and unreliable alignment regions may be frequent.

3.2.2 High-throughput sequencing, annotation errors and possible remedies

Dealing with high-throughput sequencing to conduct phylogenomic analysis introduces problems that are absent from smaller manually curated datasets.

Firstly, sequencing errors may occur, especially in the presence of homopolymers. For DNA sequences this would lead to small indel events when aligning sequences at the nucleotide level, but if they are coding sequences then their translation into amino acids will be erroneous as those events may induce artefactual frameshifts. Artefactual frameshifts can also be caused by erroneous exon boundary annotations. MACSE, an MSA software program that explicitly accounts for the underlying codon structure of protein-coding nucleotide sequences, may be used to correctly handle such coding sequences (Ranwez et al., 2011, 2018). Its unique features can help build reliable codon alignments even in the presence of (real or apparent) frameshifts.

Secondly, errors in homology annotations may lead to the inclusion of sequences erroneously considered as being homologous to others in the MSA. In this case, some MSA filtering methods may eventually be able to detect them and filter them out, but their mere

presence during the MSA process could significantly slow down the alignment process and alter the final result. If such unrelated sequences are detected, it may be worth re-aligning the remaining sequences once the rogue sequences have been removed. A similar problem arises (locally) when different splicing variants are mixed. The presence of (long fragments of) non-homologous sequences may seriously slow down and alter the alignment process. It may hence be worth trying to detect them before performing the actual alignment. This can be done using the *TrimNonHomologousFragments* subprogram of MACSE or the PREQUAL program (Whelan et al., 2018), which were specifically developed to remove long sequence fragments that are unrelated to other sequences.

A third case is related to orthology annotation errors. This more tricky case occurs when all considered sequences are homologous but some are erroneously considered as being orthologous (derived from ancestral copy by speciation) while actually being paralogous (derived from ancestral copy by duplication). When the objective is to reconstruct the species phylogeny, mixing orthologous and paralogous sequences can lead to erroneous conclusions (Chapter 2.4 [Fernández et al. 2020]). Indeed, such errors can strongly impact the phylogeny inferred since species will tend to group depending on the gene copy used to represent them rather than on their "relatedness". MSA filtering and the MACSE *TrimNonHomologousFragments* subprogram are both unhelpful in this case as the sequences are homologous and correctly aligned (residues present at the same site are homologous). However, in a phylogenomic context, and when no horizontal gene transfer is expected within the taxonomic group, such problems could be detected using the alignment of the hundreds of other genes at hand (see details in Section 3.3.6).

3.3 Principles underlying filtering methods

The underlying key ideas explained in this section are the basis of MSA filtering methods.

3.3.1 Gaps indicate hard to align and possibly saturated regions

Ultimately, sequence alignment simply consists of inserting gaps within sequences. The more gaps there are in a region, the more work the alignment method has to do, and the more likely it is that the method will generate errors. From a biological viewpoint, it is often assumed that in proteins insertions and deletions are less frequent than point substitutions. Hence, a region with multiple gaps indicates an unlikely evolutionary pattern that is most probably attributable to an MSA problem. In such regions, multiple mutations occurring at the same site are expected to be frequent and likely to obscure the phylogenetic signal.

3.3.2 Few/similar residues are expected per site

Residues within the same site are supposed to be homologous, so they are likely to share some characteristics - particularly as far as amino acid sequences are concerned. If all amino acids within a site are identical then we may be much more confident that they derive from the same ancestral amino acid than if 20 different amino acids are observed at this site. The latter case would imply not only that at least 19 substitutions have occurred to get this pattern (which could indicate a saturated site) but also that the protein has remained functional regardless of the physicochemical properties of the residue at this position. In such a situation, to filter out this part of the alignment would appear safe. Conversely, sites showing residues that share a common property – e.g., hydrophobic or positively charged – should probably be kept. Measuring residue conservation within a site can be done in

different ways, i.e. via basic measurements (number of different amino acids observed at this site, frequency of the most common amino acid) or more complex ones (measurement of site entropy, probability that two randomly picked residues are identical).

3.3.3 Models of sequence alignment

Extending the above idea, an MSA can be used to derive a model of the sequences it includes based on the observed spectrum of residues at each site. The Hidden Markov Model (HMM) provides a well-defined probabilistic model of a sequence alignment. This has many applications, such as improving homologous sequence search by BLAST-like algorithms. Of interest here is the fact that MSA HMM profile offers the opportunity to calculate a score that measures how much a given sequence fit the considered MSA. Examination of the variations of this score along the sequence can be useful to detect a potentially misaligned fragment.

3.3.4 Reliable regions are likely more robust to MSA method variations

As previously explained, MSA methods are heuristics that strive to find the MSA maximizing the chosen score (gap opening penalty, substitution matrix scores etc.). The fact that the MSA method output should be considered with caution was admirably highlighted by Landan and Graur (2007), that compare the alignment obtained with direct input DNA sequences with the reverse of the alignment obtained by aligning the reverse sequences. In a perfect world, those two alignments should be identical. In practice, they often differ, and the positions where they disagree pinpoint questionable alignment regions. These differences can be due to the presence of equally optimal scenarios at some stage of the heuristic, while the optimal scenario retained depends on the sequence orientation, and this choice impacts the next steps of the heuristic search. This heuristic search is also strongly impacted by the chosen guiding tree. Hence, tools such as Guidance (Sela et al., 2015) measure the alignment region reliability based on their stability with respect to a change in the guiding tree. Other methods go a step further and question whether the predicted residue homology relationships are stable when different penalty schemes are used (e.g. a slightly higher cost for gap opening or a different substitution matrix) or when different MSA programs are used.

3.3.5 Homologous (fragment of) sequences are expected to be similar (pre-filtering)

For most pipelines, sequence similarity is an initial criterion used to identify homologous sequences. This guarantees a minimal level of overall similarity among the sequences to be aligned. Despite this, it sometimes happen that, in a limited region of the alignment, a fragment of one (or a few) sequence does not resemble at all the rest of the alignment in this region. This sequence is thus likely not homologous to the others ones in this region, either because it was misaligned and the fragment is homologous to a distinct part of the alignment, or because it shares no homology at all with the rest of the sequences (e.g. due to alternative splicing or annotation errors). This latter situation could potentially be detected even before trying to align the sequences and a few tools have recently been developed to this effect. It is worth doing this filtering before aligning sequences since the MSA can be drastically slowed down and degraded by the presence of, particulaly, long insertions present in only one or a few sequences. The *TrimNonHomologousFragments* MACSE V2 subprogram (Ranwez et al.,

2018) and the PREQUAL program (Whelan et al., 2018) were both developed to remove fragments unrelated to other sequences, even before aligning them.

3.3.6 Orthologous sequences are supposed to be congruent over loci (post-filtering)

If an alignment contains a sequence that is not orthologous, while still being homologous, to others it is almost impossible for alignment filtering methods to detect and remove this sequence based solely on this alignment – since the problematic sequence is still highly similar to others. But in a phylogenomic context, it is possible to simultaneously consider the MSA of all considered loci to try to detect such problems. Intuitively, the distance between two sequences within a given MSA depends on the taxa from which those sequences derive (some species evolve faster than others) and on the locus represented by this MSA (some loci evolve faster than others). Having a large number of genes and taxa facilitates learning of loci and taxa evolutionary rates and hence detection of MSA having sequences significantly more distant from others than expected, considering the parameters learned using the whole set of available alignments. A simple solution to detect such non-orthologous sequences is included in the OrthoMaM v10 pipeline, while a more elaborate solution is provided by the Phylo-MCOA software package (de Vienne et al., 2012). Of course, this does not apply when the evolution of a whole gene family (orthologous + paralogous sequences) is the focus of the study.

3.4 TILI filtering methods and why they are fated to remove signals along with noise

Although some preliminary work was carried out in the early 1990s (e.g. Fernandes et al. 1993; Gatesy et al. 1993), the paper of Castresana introducing his famous Gblock software was clearly a turning point (Castresana, 2000) in the MSA literature. With more than 4,000 citations, this remains one of the most noteworthy studies in the field. Gblock defines a measurement of site conservation and basically removes any site that contains a gap and adjacent non-conserved sites, as well as stretches of non-conserved sites. Gblock is clearly a TILI-filtering method, for each site Gblock uses a series of criteria to decide whether to "take it or leave it".

Note that the number of sequences in a typical MSA has substantially increased since then, so in many applications removing any site with at least one gap is no longer a reasonable option. The latest releases of Gblock include a threshold that helps set the percentage of gaps allowed for a site. However, even in these latest versions, using default parameters will lead to the removal of any site with a gap. To really understand why removing any site with a gap is not reasonable, let us introduce an example that we will use to illustrate the inherent limitations of TILI filtering methods. Suppose you are trying to infer the phylogeny of 100 species for which you have an alignment of 5000 sites, where the sole gaps are as follows: a deletion going from sites 1 to 50 within the 1^{st} species, a deletion going from sites 51 to 100 for the 2^{nd} species, and so on up to the 100^{th} species, which has a deletion from sites 4951 to 5000. This is a great MSA matrix with only 1% of gaps, potentially conveying a strong phylogenetic signal. The Gblock default parameters, however, would remove all the sites of the alignment since they all contain at least one gap. Note that this example is intended as a criticism of the widespread blind usage of default software parameters, rather than a criticism of Gblock itself. Slightly modifying this example illustrates the problem inherent to any TILI filtering method. Suppose now that instead of gaps you have non-homologous fragments of

50 amino acids while the rest of the alignment is perfectly clean. Hence we now have, for the first 50 sites, 99 fragments of 50 amino acids that are orthologous to each other and perfectly aligned, plus a stretch of 50 unrelated amino acids (in sequence 1). Similarly, for the sites 51 to 100, we have 99 fragments of 50 amino acids that are orthologous to each other and perfectly aligned, plus a stretch of 50 unrelated amino acids in sequence 2 (see Figure 9 for an illustration of this kind of configuration). If we are using a TILI method that can only remove or keep entire sites, then there is no choice, either we keep those non-homologous fragments in the final alignment or we lose all of the sites. If a TILI-filtering approach is used for entire sequences, obviously we will end up with the same dilemma caused by the same problem. Theoretically, these non homologous fragments should generate insertions in the alignment, in practice they often don't. This problem of over alignment is well documented and hard to tackle (Katoh and Standley, 2016). It has been, for instance, emphasized on the Hedgehog-interacted protein (HHIP) to introduce the PREQUAL filtering method (Whelan et al., 2018), see https://natureecoevocommunity.nature.com/users/54859-iker-irisarri/posts/37479-automated-removal-of-non-homologous-sequence-stretches-in-phylogenomic-datasets.

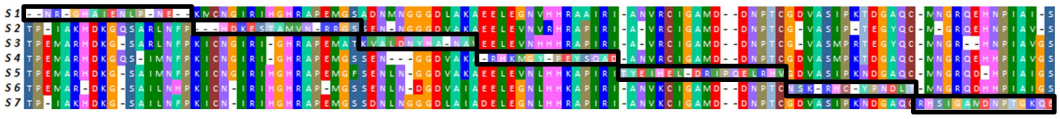


Figure 9 Schematic representation of an example of MSA where TILI-filtering methods are useless. When assuming that the alignment is perfect except in regions surrounded by the black rectangles, using a TILI approach cannot get rid of all imperfect regions without removing all the sites/sequences of the MSA. Conversely, picky methods can simply mask the residues inside those black boxes while preserving the rest of the alignment.

Note that for a TILI-filtering approach on sites, we will also end up losing the entire signal or be obliged to keep all of the noise when we have a perfect alignment of 5000 sites for 99 sequences and a last sequence completely unrelated to the others. The problem when measuring the overall conservation of a site to decide whether to keep it or not is that, as the number of sequences increases, the impact of a misaligned fragment within a sequence decreases and is masked by the conservation of the rest of the site.

Despite these limitations, TILI-filtering methods could still do a great job regarding phylogeny inference if they are able to correctly identify and remove sequences and sites containing more noise than signal. Small misaligned fragments have little impact in such cases. These limitations are much more problematic when the planned downstream analysis includes tasks such as searching for selection footprints using dN/dS or branch length analyses. In this case, every misaligned fragment has a high probability of becoming a false positive in the analysis.

4 Evaluation of MSA filtering methods

Although it would seem reasonable to rely on filtering alignment methods to obtain trustworthy alignments upon which phylogenetic analyses could rely, how to implement such automatic filtering tools is still an active yet not completely mature research field. Consequently, whether currently available automatic alignment filtering methods offer satisfactory performance or worsen the situation is still debated. Whereas each new filtering method claims to improve things, the paper of Tan, 2015, casts serious doubts on their relevance

in phylogenetic pipelines (Tan et al., 2015). However, their analysis only considered TILI filtering methods and was focused on tree topology inference. We thus decided to conduct additional tests to further evaluate the performances of TILI and picky filtering methods in a phylogenomic framework.

4.1 A benchmark of 275 genes for 116 mammal species

We performed a comparison test to evaluate the efficiency of different alignment filtering methods. We opted to use a dataset generated from the tenth release of the OrthoMaM database. OrthoMaM is a database of orthologous exon and coding sequence (CDS) alignments and phylogenetic trees. We here focus on CDS markers. The latest OrthoMaM release (v10) gathers orthologous CDS sequences from 116 fully sequenced genomes present in Ensembl and NCBI database. For each human gene, the Ensembl annotation is used to gather 1-1 orthologous sequences present in Ensembl. This core set of sequences is then enriched by searching additional 1-1 orthologous sequences within mammalian genomes only present in the NCBI database. This leads to 14,509 sets of presumably 1-1 orthologous CDS sequences containing up to 116 sequences of diverse quality. The resulting set of sequences is then processed using a dedicated pipeline to generate high quality alignments and trees.

Using OrthoMaM to build an alignment filtering benchmark has several advantages. First, the database provides not only the filtered alignment used to build each gene tree but also the raw sequences collected for each species, which is exactly what is needed to test filtering in realistic conditions. Secondly, the evolutionary history of mammals is presumably devoid of events such as genome duplication, hybridization and gene transfers between distant taxa. The tree-like evolutionary history of the 116 mammals of our data set is therefore well established (except for a few irresolutions) and most gene trees are expected to share the same topology – branch lengths, however, are expected to vary since some genes may have evolved faster than others at different periods. Phylogenomic studies have resolved the phylogeny of these 116 mammals whose topology is well known (Figure 10). This species tree, obtained with the whole OrthoMaM dataset, provides us with a "reference tree" that can be used to assess the quality of each reconstructed gene tree, after the MSA was filtered or not. To facilitate the evaluation, we focused on the 275 CDS markers of OrthoMaM where all of the 116 species are represented. In practice, for each of these 275 markers, we aligned the 116 raw (unfiltered) sequences and compared the gene tree obtained using this alignment and those obtained using different filtered versions of it, with the reference tree.

4.2 Two criteria to assess the alignment quality

To evaluate the relevance of alignment filtering methods, it is crucial to come up with a criterion according to which it is possible to measure whether a given cleaned alignment is "better" or not than the initial one.

We used two criteria to measure the quality of a filtering method based on the 275 CDS trees that are inferred using the 275 filtered alignments. The underlying assumption is that the better the tree the better the alignment used to build it. See Chapter 2.5 (Tannier et al. 2020) for another, original criterion of gene tree quality assessment.

4.2.1 Consistency of gene trees and species tree topologies

Our first proxy to estimate the quality of the gene trees obtained is to measure how congruent they are with the species tree, i.e. the reference tree. As some genes may evolve faster

than others, we focus on the tree topology, i.e. ignoring the branch lengths. More precisely, we compute the quartet distance between the reference tree and each inferred CDS tree, i.e. the smaller this distance the better. Recall that, for this taxonomic group, most genes are expected to have undergone the evolutionary history depicted by the species tree (as expected in the absence of hybridizations and lateral gene transfer events).

Briefly, to calculate the distance between two trees, all possible quartets of leaves (i.e. species) are extracted from the trees. The topology of the quartet extracted from the first tree is compared to the topology of the same quartet extracted from the reference tree. For a given quartet, their topology is either the same or different. The distance is the number, or proportion, of quartets with a different topology. We compute this distance using the tqdist program (Sand et al., 2014).

Note that, using superTriplet (Ranwez et al., 2010) to combine, into a supertree, the 275 CDS trees with 116 species provided in the OrthoMaM database, leads to the exact same tree as the species tree obtained with the 14,509 CDS of the OrthoMaM database, which also is congruent with the literature and used as a reference on the OrthoMaM website (Figure 10).

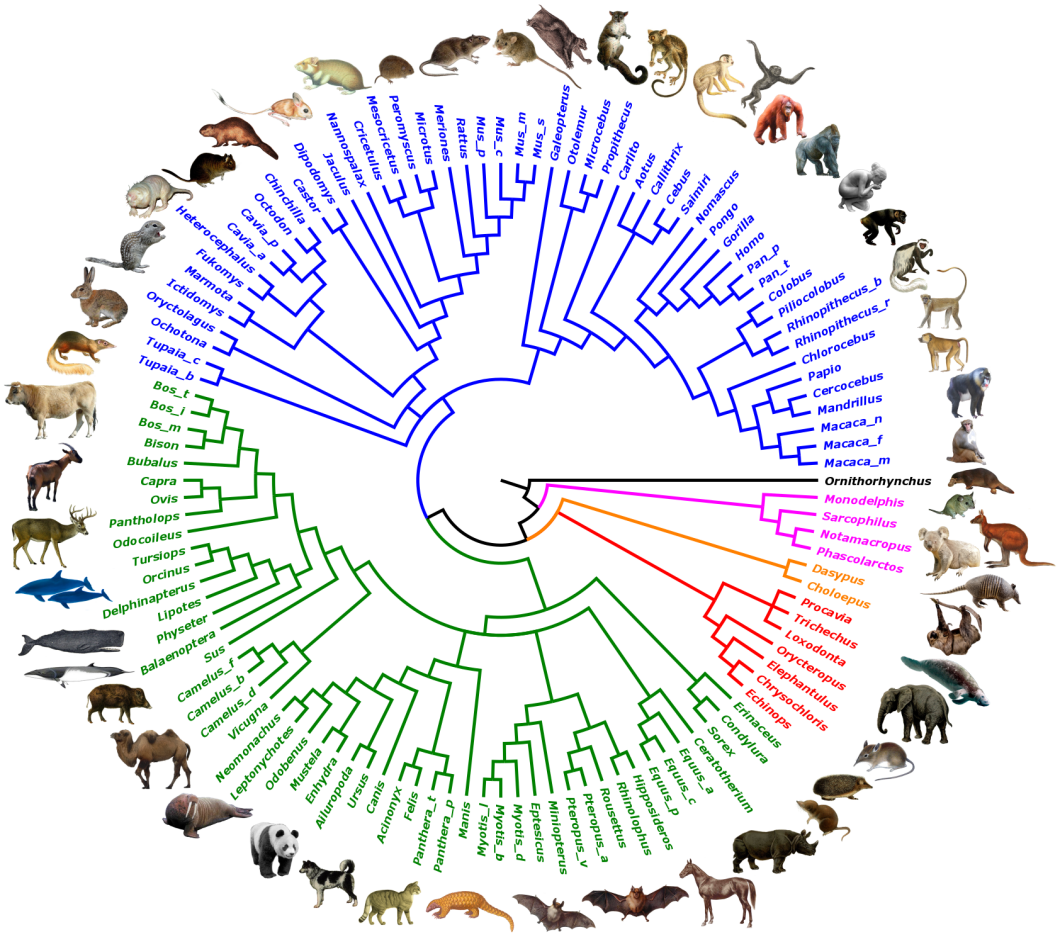


Figure 10 The OrthoMaM v10 species tree (from the OrthoMaM website). This species tree is well resolved and only a handful of irresolution remains, e.g. the clade (Procavia, Trichechus, Loxodonta) in red.

4.2.2 Consistency of terminal branch lengths among gene trees

Although useful, the first quality measurement has a major limitation, i.e. it only takes the tree topology into account, which is not the only parameter upon which misalignment may have consequences. Our second proxy to estimate the quality of gene trees aims to capture errors in branch length estimations. This is a much harder task than comparing topologies for two reasons. First, as some genes are much more constrained than others there is no reason for branch lengths to be equal across gene trees, or to those of the species tree. Secondly, as gene trees may have different topologies, there is no direct link between the branches (lengths) observed in one tree and those observed in another one (even if both have the same leaves). This latter problem can be partly overcome by focusing only on terminal branches (i.e. branches connecting species to their first parental node) as they are present in all the trees to be compared. To tackle the first problem, as we lacked a reference branch length, we used the approach described below to detect abnormally long branches.

An abnormally long terminal branch in a gene tree reflects the accumulation of private residues at one particular gene in one particular species. This can be a signature of positive selection, pseudogenization, an indication that the sequence is not orthologous to others or has been misaligned. As positive selection is assumed to be a rare phenomenon, positive selection should affect only a few sites in a sequence and not lead to very long branches. All other cases are supposed to be detected by filtering methods and could lead to very long branches (if the problem affects the whole sequence) or just a small increase in its length (if the problem affects only part of the sequence).

That said, detecting abnormally long branches cannot be done with a simple threshold regarding branch lengths. For instance, a length of 0.4 is normal for the terminal branch associated with the early-branching Ornithorhynchus, but abnormal for the terminal branch associated to Pan troglodytes, which has only recently diverged from its common ancestor to Homo sapiens. In addition, the overall evolutionary rate of the considered gene also influences the expectations, longer branches being expected in fast-evolving genes. We used a linear regression to explain the terminal branch lengths observed on inferred ML trees. The terminal branch b_{ij} leading to the species i in the gene j is viewed as $b_{ij} = \overline{b} + S_i + G_j + \epsilon_{ij}$, where $\overline{b}$ denotes the average length of all terminal branches of all gene trees; S_i is the species effect; G_j the gene effect and ϵ_{ij} the residual (the part of the branch length not captured by this model). We then considered $\epsilon^{\star}_{ij}$,the standardized residuals (a normalized version of ϵ_{ij}), as a measurement of the terminal branch length consistency. This analysis was applied independently for each filtering method.

A similar approach is used in the OrthoMaM pipeline to remove non-orthologous sequences (having a standardized residual greater than 3). The same idea has also been applied to detect positive selection by identifying (reasonably) high residuals (Wu et al., 2017). Note that the alignment filtering methods we tested here consider only one locus/gene at a time, as does the tree topology inference. Intuitively, the standardized residuals of terminal branch lengths measure the deviation of the estimation of the branch length calculated on one dataset with the estimation done on the others, while accounting for the overall gene and species evlutionary rates). The lower (the absolute value of) these standardized residuals, the more congruent the terminal branch length estimations are.

4.3 The benchmark pipeline: from orthologous sequences to a gene tree

The whole benchmark pipeline, summarized in Figure 11, is detailed in the following sections.

4.3.1 Obtaining raw and filtered alignments

For each CDS marker, the nucleotide sequences were translated into protein sequences. The protein sequences were aligned with MAFFT and the nucleotide alignment was derived from the protein alignment using the back-aligned procedure provided by the 'egglib' python library (De Mita and Siol, 2012), see Figure11 A. Then, seven MSA filtering methods were applied on the 275 nucleotide raw alignments (Figure11 B).

Figure 11 Schematic representation of the different filtering processes. Part A) depicts the alignment process: the nucleotide coding sequences are translated into amino acid sequences that are then aligned with MAFFT. The resulting amino acid alignment is then used to derive the nucleotide alignment. Part B) depicts the seven filtering processes that we compare to the "no filtering" approach.

Note that when CDS are the raw material for evolutionary study, the best strategy is to align them based on their translated sequences since amino acid sequences are more similar than their nucleotide counterparts (due to the genetic code redundancy and selective pressure acting at the amino acid level) and are hence easier to align. Once the protein alignments are obtained and cleaned, they may be used to derive the corresponding nucleotide alignments. Alignments are facilitated by using translated amino acid sequences. Filtering at the amino acid level is also easier thanks to the richer amino acid alphabet while ensuring that the codon structure is preserved. Whether inferring the phylogeny based on the (filtered) amino acid alignments or based on the (filtered) nucleotide alignments derived from them depends on the studied taxonomic level. Protein alignments will be better for deep phylogeny inferences (avoiding nucleotide saturation) while nucleotide alignments will be better for recent phylogenies (allowing observation of sequence differences that are masked at the amino acid level). According to what is done in the OrthoMaM pipeline, here we opted to infer gene trees based on the nucleotide alignments.

The filtering methods tested are presented in Table 1 (columns 1-3) and in Figure 11. The options used are detailed below.

- "no filtering": the nucleotide coding sequences are translated into amino acid sequences that are then aligned with MAFFT. The resulting amino acid alignment is then used to derive the nucleotide alignment, which serves as a reference point.
- Gblock is used to filter reference alignments with the codon option activated (option `-t=c`) and the possibility of having gaps (option `-b5=h`, i.e. only sites where gaps are present in less than half of the sequences could be kept).
- trimAl does not provide a codon filtering option, but its amino acid filtering provides, as

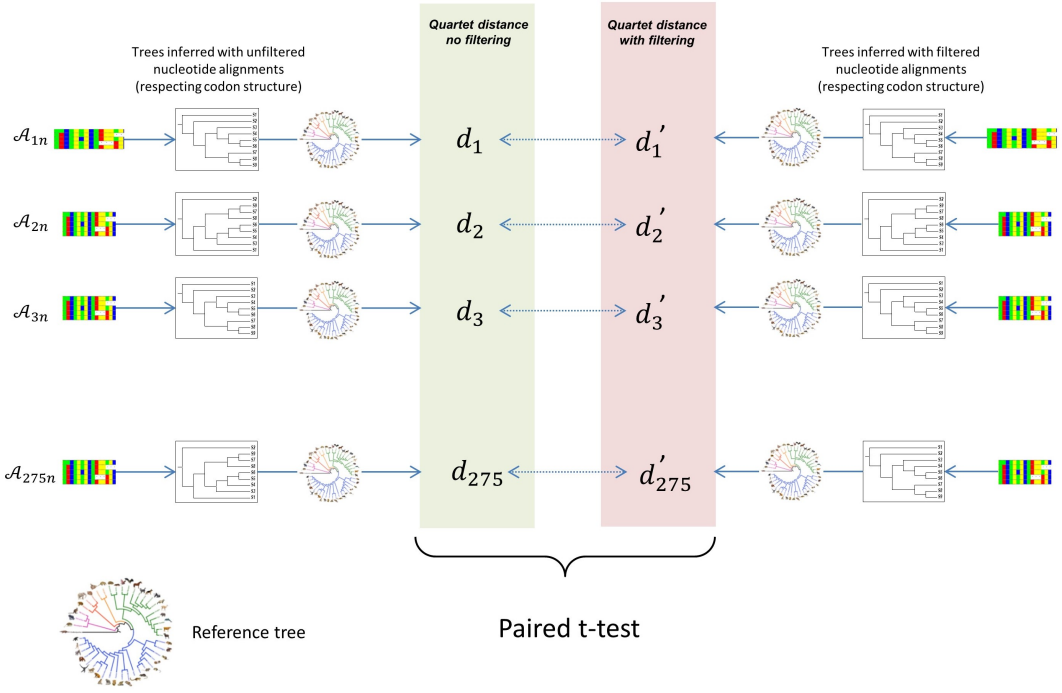


Figure 12 Schematic representation of the test used to assess the impact of a given alignment filtering method on gene tree topologies.

tested methods			% of removed nucleotides		normalized q-distance	
name	ref	category	avg	median	avg	p-value
no filtering	-	-	0	0	0.129	-
Gblock	(Castresana, 2000)	TILI	6.07	3.92	0.129	0.88
trimAl	(Capella-Gutierrez et al., 2009)	TILI	3.76	1.36	0.128	0.75
BMGE	(Criscuolo and Gribaldo, 2010)	TILI	4.77	3.11	0.127	0.23
Guidance_TILI	(Penn et al., 2010; Sela et al., 2015)	TILI	4.95	3.13	0.129	0.97
Guidance_picky	(Penn et al., 2010; Sela et al., 2015)	picky	2.90	1.99	0.128	0.59
OMM_MACSE	(Scornavacca et al., 2019)	picky	2.20	1.58	0.121	3.2E-3
HmmCleaner	(Di Franco et al., 2019)	picky	2.22	1.55	0.120	1.4E-5

Table 1 Impact of filtering methods on the gene tree topologies. The first three columns provide information regarding each tested filtering method, its name, the corresponding paper(s) and the type of filtering applied. We computed the percentage of nucleotides removed from each filtered alignment; this led to 275 values per method whose average and median are provided in columns 4 and 5, respectively. We computed the normalized quartet distances between the reference tree and each inferred tree; this led to 275 values per method whose average is provided in column 6. For a given filtering method, the 275 normalized quartet distances are compared (using a paired t-test, see Figure 12) with the 275 normalized quartet distances obtained with "no filtering"; the p-value of this test is provided in the last column.

output, indices of the removed amino acid sites; we used this information to derive the corresponding filtered nucleotide alignment.

- BMGE can handle coding nucleotide alignments with its codon option activated (option `-t CODON` for BMGE).
- Guidance works directly on the unaligned coding nucleotide sequences as it includes an alignment step able to handle amino acid translation and back translation (options `--msaProgram MAFFT --seqType codon`). By default Guidance uses a TILI strategy.
- maskLowScoreResidues.pl is a handy perl script that comes with Guidance. It takes advantage of Guidance output files to filter the initial alignment at the residue level (picky approach). As the script provides no default threshold, we used a threshold of 0.9 (option `0.9 nuc`), as also carried out in the paper mentioned on the Guidance documentation of this feature (Privman et al., 2012).
- OMM_MACSE is used with default options. Starting with unaligned nucleotide coding sequences, this pipeline chains sequence pre-filtering (removing long non-homologous fragments) and detection of frameshifts (using MACSE), alignment of amino acid sequences (using MAFFT), filtering of the resulting alignment (using HmmCleaner) and post-filtering (using MACSE) to mask isolated residues (those surrounded by gaps after HmmCleaner filtering) and to completely remove sequences for which more than half of the residues were masked during the pipeline.
- HmmCleaner works on amino acid alignments and the release we used generated a filtered alignment where masked residues were replaced by a specific user defined character. We used the '$' symbol to mask residues and a threshold of 10 (options `--del-char '\$' 10`). The *reportMaskAA2NT* subprogram of MACSE was then used to derive, thanks to the positions of the '$' symbols, the filtered alignment of the corresponding nucleotide coding sequences.

4.3.2 Obtaining phylogenetic trees and quartet distances

We have eight alignments for each of the 275 CDS markers: one raw alignment and seven filtered ones. For each of these 8×275 alignments, a maximum likelihood tree was inferred with phyML under a GTR+gamma model (Guindon et al., 2010) and the (normalized) quartet distance between the inferred tree and the reference tree was computed using tqdist (Sand et al., 2014). Note that since tqdist can only compute the quartet distance between two trees having exactly the same set of leaves, this criterion is not applicable to the few cases in which OMM_MACSE removes an entire sequence from the alignment. For this filtering approach, the total number of quartets with different topologies is hence not directly comparable to that of other filtering methods and only the average standardized quartet distance is comparable.

4.4 Results

4.4.1 Some filtering methods improve gene tree topologies much more than others

The quartet distances observed between the gene trees obtained and the reference tree are summarized in Table 2. All filtering methods lead to gene trees whose topology is more similar to that of the species tree than those obtained without filtering. Indeed, the normalized quartet distances to the species tree are about 0.120 for HmmCleaner, 0.121 for OMM_MACSE, 0.127 for BMGE, 0.128 for trimAl and the picky version of Guidance, 0.129 without filtering or with Gblock and the TILI version of Guidance. This gene tree topology improvement is not significant, however (based on a paired t-test 5%), with the exception of two filtering methods: HmmCleaner and the OMM_MACSE pipeline (that also relies on HmmCleaner). These findings are in line with the idea that picky filtering methods, which are able to mask individual residues, are more powerful than TILI methods, which can only remove entire sites or sequences, as explained above (Section 3.4). Note that the picky version of Guidance (normalized quartet distance 0.128) is only slightly better than its TILI counterpart (0.129) and that it does not improve the gene tree topologies as much as HmmCleaner(0.120).

4.4.2 Filtering methods have a notable impact on terminal branch lengths

The fourth column of Table 2 provides the count of gene tree terminal branch lengths longer than 0.5 (on average one mutation every two sites along these branches), excluding long branches leading to the Ornithorhynchus outgroup. Such long branches are highly unlikely and generally result from the presence of non-homologous or misaligned sequences. With just 4 such branches, OMM_MACSE is the most efficient method for this criterion, closely followed by HmmCleaner which has 7 such branches. All other methods lead to more than 20 of these long branches. Guidance_picky seems especially inefficient with 60 of these long branches while there are only 30 without any filtering. This seems to at least partially be related to the fact that this method sometimes masks all residues of a sequence without removing it from the alignment, hence the position of this sequence in the gene tree, as well as its branch length, are completely meaningless. Such extreme cases are correctly handled by OMM_MACSE which removes the sequence from the alignment when too many residues are masked, but they do not seem to be handled by other methods, which may mask all, or almost all, of the residues in a sequence without warning the end user.

tested methods			number of abnormally long branches	
name	ref	category	length > 0.5 excluding Ornithorhynchus	$\epsilon^{\star}_{ij} > 3$ and length > 0.01
no filtering	-	-	30	152
Gblock	(Castresana, 2000)	TILI	25	153
trimAl	(Capella-Gutierrez et al., 2009)	TILI	22	145
BMGE	(Criscuolo and Gribaldo, 2010)	TILI	23	157
Guidance_TILI	(Penn et al., 2010; Sela et al., 2015)	TILI	22	150
Guidance_picky	(Penn et al., 2010; Sela et al., 2015)	picky	60	85
OMM_MACSE	(Scornavacca et al., 2019)	picky	4	27
HmmCleaner	(Di Franco et al., 2019)	picky	7	31

Table 2 Impact of filtering methods on terminal branch lengths of gene trees. The first three columns provide information regarding each tested filtering method, its name, the corresponding paper(s) and the type of filtering applied. Column four indicates the number of terminal branches longer than 0.5, which usually reveals an alignment problem except, occasionally for the outgroup taxon. Column five indicates the number of abnormally long branches according to a linear model using species and genes as fixed parameters. See main text for details.

The count of abnormally long branches detected using a more elaborate test (standardized residuals > 3 and branch length > 0.01) that takes the fact that the evolutionary rates depend on the considered species and gene, leads to the same overall results. As detailed in the fifth column of Table 2, trees produced using the OMM_MACSE pipeline and HmmCleaner have many fewer abnormally long branches. These two methods have about 30 problematic branches detected, while 145 to 157 such problematic branches are detected with Gblock, Guidance_TILI, BMGE, Gblock, and trimAl or in the absence of filtering (152 problematic branches). The only discrepancy between these two abnormally long branch measurements concerns Guidance. While Guidance_picky has many fewer abnormally long branches according to the residual errors than Guidance_TILI (85 vs 150), it has many more branches longer than 0.5 (150 vs 22).

Figure 13 below illustrates one such case of abnormally long branches, where the part of the sequence removed by OMM_MACSE and HmmCleaner is obviously not orthologous to the others.

The abnormal branches detected using the above tests are extreme cases where a large portion of a sequence should be masked by filtering. When only a relatively short fragment of a sequence is problematic, and should be masked, its presence may not have sufficient impact on the corresponding branch length or its standardized residual to make any difference when using these tests. Figure 14 highlights the overall impact of HmmCleaner filtering on terminal branch lengths. In this figure, we plotted the length of each terminal branch estimated with HmmCleaner filtering (Y-axis) and without filtering (X-axis). Obviously, the main trend is that most of the points are below the y=x line, clearly indicating that using HmmCleaner tends to reduce terminal branch lengths. The same trend is also observed with OMM_MACSE.

5 Conclusion

This chapter details the key steps of MSA with an emphasis on the underlying arbitrary choices (e.g. gap opening costs) and inherent limitations (e.g. heuristic searches to optimize the alignment score). Given these limitations, it would seem natural to try to post-process alignments and filter out the erroneous parts. Though some obvious errors are easily spotted by the expert human eye, defining criteria to automatically determine which parts of the alignment are correct and incorrect is a difficult task. In recent years, several alignment filtering methods have been proposed based on criteria that we informally explain here. We believe that, independently of the filtering criteria used, the main distinction between these filtering methods concerns whether or not they can remove part of a site (or sequence) by simply masking some residues of a site (or sequence). We have called those that can "picky" and those that cannot "TILI". Using a simple example, we explained why TILI methods are inherently less able to spot errors in large alignments.

We assessed the impact of alignment filtering methods by comparing their impact on the gene tree inference of 275 CDS datasets, each consisting of 116 putative orthologous mammal sequences. As expected, the picky HmmCleaner method, either used alone or included as part of the OMM_MACSE pipeline, performed best. HmmCleaner and OMM_MACSE had a positive impact on the topology of individual gene trees. Using one of those two filtering approaches led to gene tree topologies significantly closer to that of the species tree, compared to topologies obtained in the absence of alignment filtering. The other tested

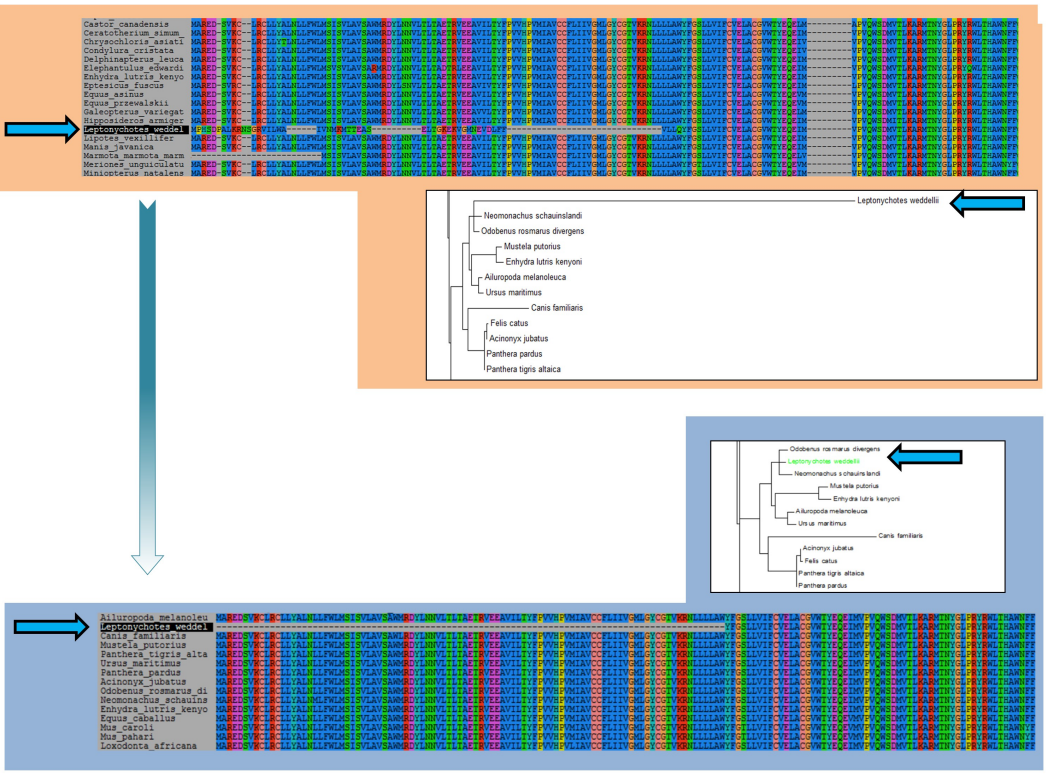


Figure 13 A case of an abnormally long branch which has been corrected by OMM_MACSE and HmmCleaner but cannot be fixed by a PICKY filtering method.

filtering methods had no significant impact on gene tree topologies. Alignment filtering also had a major impact on branch lengths, which is an often overlooked aspect. Here again, HmmCleaner and OMM_MACSE yielded the best results in our tests, with a significant reduction in the number of abnormally long terminal branches and an overall increase in the consistency of the terminal branch lengths among the independently inferred gene trees. Note that those two methods are closely linked. Indeed OMM_MACSE is a coding sequence alignment dedicated pipeline that includes an HmmCleaner filtering step together with some specific pre- (before alignment) and post- (after HmmCleaner) cleaning steps.

Our results do not contradict those of (Tan et al., 2015). Indeed, these authors did not observe any benefit of using TILI filtering methods regarding the gene tree topologies, nor did we. We do not believe that this indicates that the filtering alignment is counter-productive for phylogeny inference, but it simply emphasizes the limitations inherent to TILI filtering methods. Our study also included some picky filtering methods and we were able to illustrate

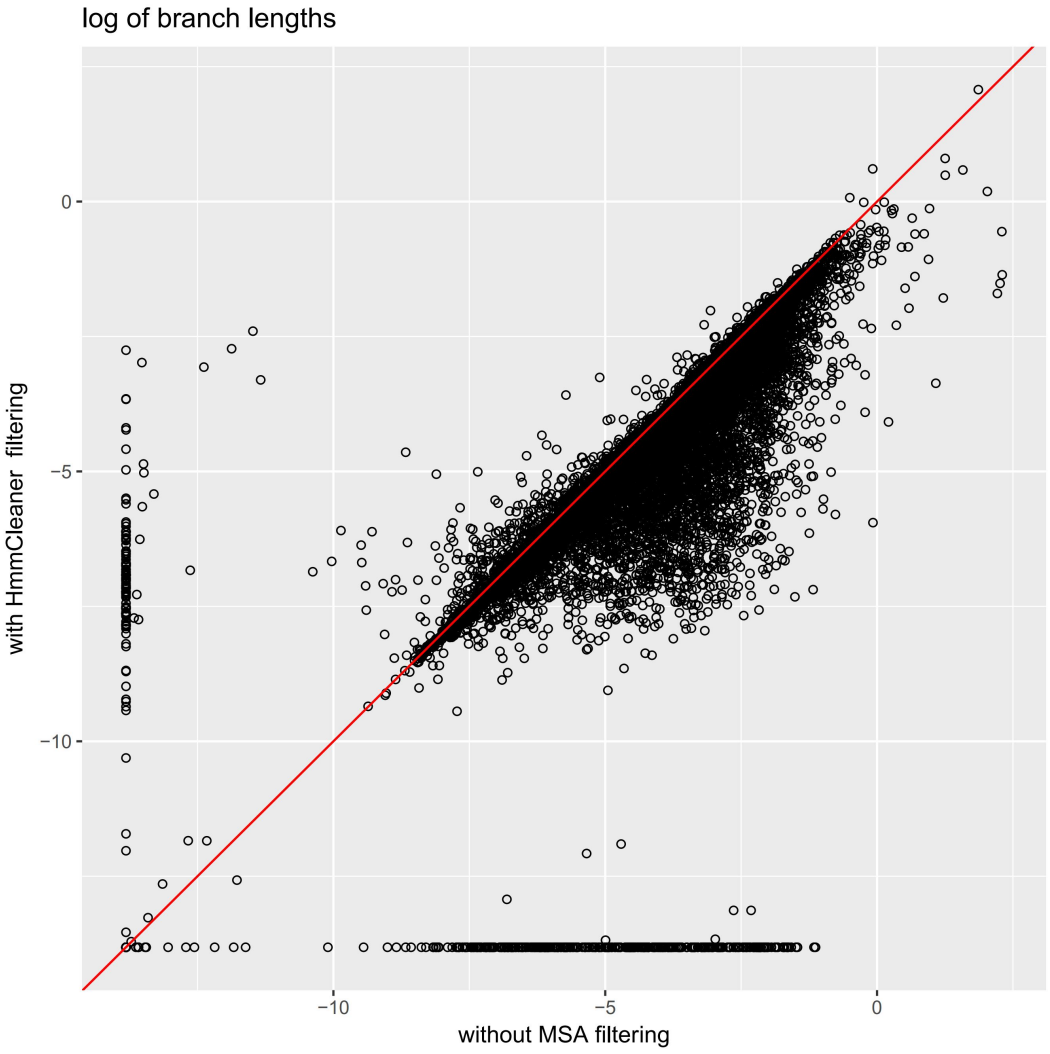


Figure 14 Dot plots comparing the log of terminal branch lengths obtained without MSA filtering (x-axis) and with HmmCleaner filtering (y-axis). The $y = x$ line is drawn in red to facilitate interpretation.

the benefits of these. Moreover, it is an over simplification to consider phylogenetic inference only in terms of gene tree topologies. Here we have shown that alignment filtering may also have a marked impact on branch length estimations. Similarly, it would be worth investigating their impact on branch supports (bootstrap or posteriors), on the estimates of other evolutionary model parameters (e.g. substitution rates), and on downstream analyses – e.g. positive selection detection.

It should be noted that the potential effects (positive or negative) of MSA filtering methods also depend on the initial alignment quality. In other words, if great care is taken to ensure that the sequences to be aligned are homologous and do not contain long stretches of non-homologous residues, filtering methods will probably fail to substantially improve the MSA, and might even worsen it. However, when automatic homology search methods are applied in a phylogenomic project, applying picky-filtering methods is worthwhile.

References

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial dna. *J Mol Evol*, 42(4):459–68.

Altschul, S. F. (1989). Gap costs for multiple sequence alignment. *J Theor Biol*, 138(3):297–309.

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–3.

Cartwright, R. A. (2006). Logarithmic gap costs decrease alignment accuracy. *BMC Bioinformatics*, 7:527.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*, 17(4):540–52.

Criscuolo, A. and Gribaldo, S. (2010). Bmge (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol*, 10:210.

De Mita, S. and Siol, M. (2012). Egglib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet*, 13:27.

de Vienne, D. M., Ollier, S., and Aguileta, G. (2012). Phylo-mcoa: a fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Mol Biol Evol*, 29(6):1587–98.

Di Franco, A., Poujol, R., Baurain, D., and Philippe, H. (2019). Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *Bmc Evolutionary Biology*, 19.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7.

Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25(4):351–60.

Fernandes, A. P., Nelson, K., and Beverley, S. M. (1993). Evolution of nuclear ribosomal rnas in kinetoplastid protozoa: perspectives on the age and origins of parasitism. *Proc Natl Acad Sci U S A*, 90(24):11608–12.

Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.

Fleissner, R., Metzler, D., and von Haeseler, A. (2005). Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst Biol*, 54(4):548–61.

Gatesy, J., DeSalle, R., and Wheeler, W. (1993). Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol Phylogenet Evol*, 2(2):152–7.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst Biol*, 59(3):307–21.

Herman, J. L., Challis, C. J., Novak, A., Hein, J., and Schmidler, S. C. (2014). Simultaneous bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Mol Biol Evol*, 31(9):2251–66.

Katoh, K. and Standley, D. M. (2016). A simple method to control over-alignment in the mafft multiple sequence alignment program. *Bioinformatics (Oxford, England)*, 32(13):1933–1942.

Landan, G. and Graur, D. (2007). Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*, 24(6):1380–3.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, 21(6):1095–109.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol*, 25(7):1307–20.

Lowe, C. and Rodrigue, N. (2020). Detecting adaptation from multi-species protein-coding dna sequence alignments alignments. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.5, pages 4.5:1–4.5:18. No commercial publisher | Authors open access book.

Loytynoja, A. and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5883):1632–5.

Lunter, G., Miklos, I., Drummond, A., Jensen, J. L., and Hein, J. (2005). Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, 6:83.

Madhusudhan, M. S., Marti-Renom, M. A., Sanchez, R., and Sali, A. (2006). Variable gap penalty for protein sequence-structure alignment. *Protein Eng Des Sel*, 19(3):129–33.

Morrison, D. A. (2006). Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany*, 19(6):479–539.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–17.

Ogden, T. H. and Rosenberg, M. S. (2007). Alignment and topological accuracy of the direct optimization approach via poy and traditional phylogenetics via clustalw + paup*. *Syst Biol*, 56(2):182–93.

Penn, O., Privman, E., Landan, G., Graur, D., and Pupko, T. (2010). An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol*, 27(8):1759–67.

Privman, E., Penn, O., and Pupko, T. (2012). Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol Biol Evol*, 29(1):1–5.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Ranwez, V. (2016). Two simple and efficient algorithms to compute the sp-score objective function of a multiple sequence alignment. *Plos One*, 11(8).

Ranwez, V., Criscuolo, A., and Douzery, E. J. P. (2010). Supertriplets: a triplet-based supertree approach to phylogenomics. *Bioinformatics*, 26(12):i115–i123.

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., and Delsuc, F. (2018). Macse v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol Biol Evol*, 35(10):2582–2584.

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. (2011). Macse: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One*, 6(9):e22594.

Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., and Pedersen, C. N. (2014). tqdist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*, 30(14):2079–80.

Scornavacca, C., Belkhir, K., Lopez, J., Dernat, R., Delsuc, F., Douzery, E. J. P., and Ranwez, V. (2019). Orthomam v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol Biol Evol*, 36(4):861–862.

Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. (2015). Guidance2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res*, 43(W1):W7–14.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., and Dessimoz, C. (2015). Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Syst Biol*, 64(5):778–91.

Tannier, E., Bazin, A., Davín, A. A., Guéguen, L., Bérard, S., and Chauve, C. (2020). Ancestral genome organization as a diagnosis tool for phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.5, pages 2.5:1–2.5:19. No commercial publisher | Authors open access book.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80.

Thompson, J. D., Linard, B., Lecompte, O., and Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One*, 6(3):e18093.

Thorne, J. L. and Kishino, H. (1992). Freeing phylogenies from artifacts of alignment. *Mol Biol Evol*, 9(6):1148–62.

Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *J Comput Biol*, 1(4):337–48.

Wheeler, T. J. and Kececioglu, J. D. (2007). Multiple alignment by aligning alignments. *Bioinformatics*, 23(13):i559–68.

Wheeler, W. (1996). Optimization alignment: The end of multiple sequence alignment in phylogenetics? *Cladistics*, 12(1):1–9.

Wheeler, W. C. (2003). Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search. *Cladistics*, 19(3):261–8.

Whelan, S., Irisarri, I., and Burki, F. (2018). Prequal: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics*, 34(22):3929–3930.

Wu, J., Yonezawa, T., and Kishino, H. (2017). Rates of molecular evolution suggest natural history of life history traits and a post-k-pg nocturnal bottleneck of placentals. *Current Biology*, 27(19):3025 – 3033.e5.

Zou, Z. and Zhang, J. (2020). The nature and phylogenomic impact of sequence convergence. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.6, pages 4.6:1–4.6:17. No commercial publisher | Authors open access book.

Chapter 2.3 Accurate alignment of (meta)barcoding data sets using MACSE

Frédéric Delsuc[1]
Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France
frederic.delsuc@umontpellier.fr
https://orcid.org/0000-0002-6501-6287

Vincent Ranwez[2]
AGAP, Université de Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France
vincent.ranwez@supagro.fr
https://orcid.org/0000-0002-9308-7541

Abstract

Twenty years of standardized DNA barcoding practice have resulted in millions of sequences being produced for a handful of molecular markers in a wide range of fungi, animal and plant species. Despite some basic quality controls, reference barcoding data sets deposited in the Barcode of Life Datasystem (BOLD) database are not immune to sequencing errors and undetected pseudogenes. Such database inaccuracies can significantly bias subsequent species delimitation and biodiversity estimation based on DNA barcoding. These potential problems are amplified in metabarcoding studies containing thousands of sequences produced using high throughput sequencing technologies. Here, we propose a pipeline based on MACSE v2, an extended version of our codon-aware multiple sequence alignment software accounting for frameshifts and stop codons. The MACSE_BARCODE pipeline allows the accurate alignment of hundreds of thousands of protein-coding barcode sequences. Re-analyses of published data sets confirm that MACSE v2 is able to automatically detect most sequencing errors previously identified manually. The proposed alignment strategy hence alleviates the risk of incorrect species delimitation due the incorporation of sequencing errors or undetected pseudogenes. By applying the MACSE_BARCODE pipeline to mammal, ant, and flowering plant barcode sequences available in BOLD, we highlight several cases of database errors and provide curated reference alignments for the main protein-coding barcode genes. We anticipate our approach to be particularly useful for metabarcoding studies in which thousands of new sequences need to be compared to a reference database for subsequent taxonomic assignment. This might prove particularly helpful for diet characterization studies and large-scale biodiversity assessments through environmental DNA metabarcoding. The new MACSE_BARCODE pipeline is distributed as Nextflow workflows that are available from the MACSE project webpage (`https://bioweb.supagro.inra.fr/macse/`).

How to cite: Frédéric Delsuc and Vincent Ranwez (2020). Accurate alignment of (meta)barcoding data sets using MACSE. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 2.3, pp. 2.3:1–2.3:31. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

[1] FD was funded by the European Research Council via the ERC-2015-CoG-683257 ConvergeAnt project.
[2] VR was supported by the CIRAD UMR AGAP HPC Data Center of the South Green Bioinformatics platform (`http://www.southgreen.fr/`).

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 2.3; pp. 2.3:1–2.3:31

A book completely handled by researchers.
No publisher has been paid.

1 DNA (meta)barcoding and MACSE

Since the birth of molecular systematics in the mid 1960s (Zuckerkandl and Pauling, 1965; Sarich and Wilson, 1967) evolutionary biologists have used molecules to characterize species biodiversity and evolution. The concept of using DNA sequences to distinguish species and reconstruct their phylogenetic relationships based on a universal molecular marker has been adopted early on by microbiologists following the pioneering work of Carl Woese and collaborators on the *16S ribosomal RNA* (*16S rRNA*) gene (Woese and Fox, 1977; Woese et al., 1990). The proposal of using a few standardized universal molecular markers for species identification via the so-called "DNA barcoding" approach has been later popularized by Hebert et al. (2003a). This approach has since been largely embraced by the molecular ecology community and has found many applications from large-scale species inventories (Hebert et al., 2004; Smith et al., 2005; Ward et al., 2005; Lahaye et al., 2008), to more global biodiversity assessments through metabarcoding thanks to the developments of high-throughput sequencing of environmental DNA (Taberlet et al., 2012; Ji et al., 2013; Bohmann et al., 2014).

In practice, almost two decades of DNA barcoding has resulted in the build-up of reference sequence databases of universal barcoding markers linked to biological specimens for fungi, animals, and plants. The Barcode of Life Datasystem (BOLD, Ratnasingham and Hebert 2007) version 4 now contains more than 8 million barcode sequences representing more than 300,000 species. The vast majority of these sequences are from the mitochondrial *cytochrome c oxidase I* (*COI*, Hebert et al. 2003a) gene for animals and from the chloroplastic *Ribulose-1,5-bisphosphate carboxylase/oxygenase* (*rbcL*, Kress and Erickson 2007) and *Maturase K* (*matK*, Dunning and Savolainen 2010) genes for plants. These protein-coding genes have been selected for their ability to discriminate species by showing a clear separation between intraspecific polymorphism and interspecific divergence (Hebert et al., 2003b; CBOL Plant Working Group, 2009). These reference databases offer a great resource and are routinely used to assess the taxonomic assignment of newly produced barcoding sequences in an ever-growing number of barcoding and metabarcoding projects. In light of their importance to the field, quality controls have been introduced for sequence deposit with recommendations on how to validate sequences to be included in the database. Despite these precautions, the BOLD database is not immune to sequencing errors leading to bad sequence quality and undetected pseudogenes (Stoeckle and Kerr, 2012).

One particular problem is the potential integration of nuclear copies of mitochondrial derived genes (*numts*, Lopez et al. 1994) in *COI*, which are known to be widespread in a number of animal groups (Bensasson et al., 2001). Practical solutions have been proposed for limiting the co-amplification of *numts* with mitochondrial barcoding markers (Moulton et al., 2010; Calvignac et al., 2011) but *numts* are difficult to detect in practice and they create obvious problems for species delimitation (Song et al., 2008). For protein-coding barcode genes, alignment to reference sequences, and detection of frameshifts and stop codons are part of the requirements for sequence deposition in BOLD. However, it has been shown that errors in barcoding sequences tend to be clustered at sequence extremities; this illustrates the potential problem of relying only on stop codon detection based on a short fragment within which frameshifts close to extremities will go undetected (Stoeckle and Kerr, 2012). Buhay (2009) highlighted the problems of quality control by examining sequences visually, but the huge number of new DNA barcoding sequences being produced –coupled with the development of metabarcoding– make visual detection and manual correction impossible. Voices have even been raised to question the ability of the DNA barcoding community to

embrace high-throughput sequencing (Taylor and Harris, 2012) while others have underlined the bioinformatic challenges associated with the rise of DNA metabarcoding (Coissac et al., 2012).

In this context, our multiple sequence alignment program MACSE (Ranwez et al., 2011), which accounts from frameshifts and stop codons, has been adopted early on by the DNA barcoding community. MACSE is indeed particularly suited to deal with protein-coding barcode sequences that are generally highly conserved at the protein level. Consequently, the first version of MACSE v1 was quickly introduced into metabarcoding bioinformatic pipelines as a denoising step, allowing the automatic detection of stop codons and frameshifts in sequences produced by error-prone sequencing technologies –such as 454 Life Sciences pyrosequencing (Yu et al., 2012; Ji et al., 2013; Ramirez-Gonzalez et al., 2013; Yang et al., 2014). This problem was later alleviated by the development of metabarcoding protocols based on Illumina short read sequencing (Liu et al., 2013). However, third generation long-read sequencing technologies, e.g. Pacific Biosciences and Oxford Nanopore, still suffer from relatively high error rates linked to homopolymers. Besides sequencing error detection, MACSE is also relevant as a tool to automatically spot *numts* and pseudogenes, all the more so as protein-coding markers present numerous advantages as barcoding markers compared to non-coding ones (Andújar et al., 2018).

MACSE is, however, much more than a protein-coding sequence denoising tool. It is first and foremost a multiple sequence alignment program that has recently been enriched with a toolkit of subprograms to handle protein-coding alignments (MACSE v2, Ranwez et al. 2018). A DNA barcoding application for which MACSE has been and could be particularly useful is diet assessment via metabarcoding (Pompanon et al., 2012). This kind of study typically requires thousands of newly produced barcoding sequences to be compared against a reference database such as BOLD to perform taxonomic assignment of prey items. In this case, as in most other metabarcoding applications (Leray and Knowlton, 2015), accurately aligning the newly produced sequences to the sequences of the reference database could be particularly valuable. MACSE has effectively been used for doing so in the context of the Moorea Biocode project, which aimed to create the first comprehensive inventory of all non-microbial life in a complex tropical ecosystem (Leray et al., 2013). Indeed, working from a multiple sequence alignment allows to leverage the power of phylogenetics for taxonomic assignment instead of relying on simple sequence similarity searches. The advantages of using alignments and phylogenetic trees have long been recognized in the microbiome field, in which the main dedicated pipelines based on *16S rRNA* metabarcoding, such as MOTHUR (Schloss et al., 2009) and QIIME (Caporaso et al., 2010), are routinely used to performed taxonomic assignment based on curated databases of sequence alignments and phylogenetic trees with updated taxonomy such as Greengenes (DeSantis et al., 2006) and SILVA (Pruesse et al., 2007). Ultimately, with the availability of reference alignments and phylogenetic trees, taxonomic assignment could be based on probabilistic evolutionary placement as implemented in programs such as pplacer (Matsen et al., 2010), RAxML_EPA (Berger et al., 2011), and RAPPAS (Linard et al., 2019). Unfortunately, it is far from being the case in the DNA barcoding field, as no such reference alignments and trees exist. Indeed, if public sequence data can easily be downloaded from the BOLD database for the different taxonomic groups represented, the resulting files contain unaligned sequences consisting of a mix of different barcoding fragments and markers.

In this chapter, after illustrating the usefulness of MACSE v2 to deal with sequencing errors and *numts* in barcoding data sets, we present a new pipeline named MACSE_BARCODE, which aims at accurately aligning hundreds of thousands of protein-coding barcode sequences.

By applying MACSE_BARCODE to mammal, ant, and flowering plant sequences of the BOLD database, we provide high quality reference alignments of *COI* (mammals and ants), and *rbcL* and *matK* (flowering plants) as freely available resources for the DNA barcoding community.

2 Using MACSE to align protein-coding (meta)barcoding data sets

In this section, we provide a short overview of the features included in MACSE v2, with an emphasis on the subprograms and options that are particularly useful when dealing with protein-coding barcoding data. The MACSE_BARCODE pipeline is based on these subprograms and could be run in two command lines without *a priori* knowledge of the underlying details. However, getting familiar with the subprograms at the core of the barcoding pipeline will allow the user to understand what is in the blackbox and, thus, help with the interpretation of the results. A more detailed presentation of MACSE v2 can be found in Ranwez et al. (2020).

2.1 A quick overview of MACSE v2

MACSE v1 (Ranwez et al., 2011) was originally developed as a single program to align nucleotide sequences of protein-coding genes while accounting for frameshifts and stop codons. Its second version (MACSE v2, Ranwez et al. 2018) extended the initial release with a suite of subprograms and a Graphical User Interface (GUI). The main subprogram is *alignSequences*, which is at the core of MACSE. The other subprograms constitute a rich toolkit for handling alignments of protein-coding sequences. The current MACSE version v2.03 includes 14 subprograms that, for instance, allow: (1) refining an existing alignment (*refineAlignment*), (2) trimming an existing alignment (*trimAlignment*), (3) removing non-homologous sequence fragments (*trimNonhomologousFragments*), (4) excluding nucleotides, codons, or sites involved in frameshifts and stop codons (*exportAlignment*), (5) merging two distinct alignments (*alignTwoProfiles*), and (6) enriching an alignment by sequentially adding sequences (*enrichAlignment*). These subprograms can thus be combined or sequentially executed to design simple alignment pipelines. All analyses in this chapter were performed using the command line version of the latest MACSE release (but note that the analyses that do not involve the metabarcoding pipeline could also be reproduced using the GUI, see Figure 1).

MACSE can be freely downloaded as a single jar file from its dedicated website (`https://bioweb.supagro.inra.fr/macse/`). As it is written in the JAVA language, MACSE does not need installation and should run under any operating system (Windows, MAC OS, Linux), provided that the Java Runtime Environment (JRE) is installed on the computer. The MACSE website contains detailed documentation and tutorials presenting several examples of applications. Once downloaded, MACSE can be directly launched by double clicking on the `macse_v2.03.jar` file or by typing the following command in a terminal:

```
$ java -jar ./macse_v2.03.jar
```

In both cases, this should open the GUI version of MACSE (Figure 1). The “Programs” menu allows choosing from the different subprograms that were designed to align and manipulate protein-coding sequence `.fasta` files. Once a subprogram is selected, a brief description of this subprogram is provided and the different options are available from the different tabs corresponding to mandatory options, e.g. output file names, alignment parameters, etc... When an option field is selected by clicking on it, the related documentation

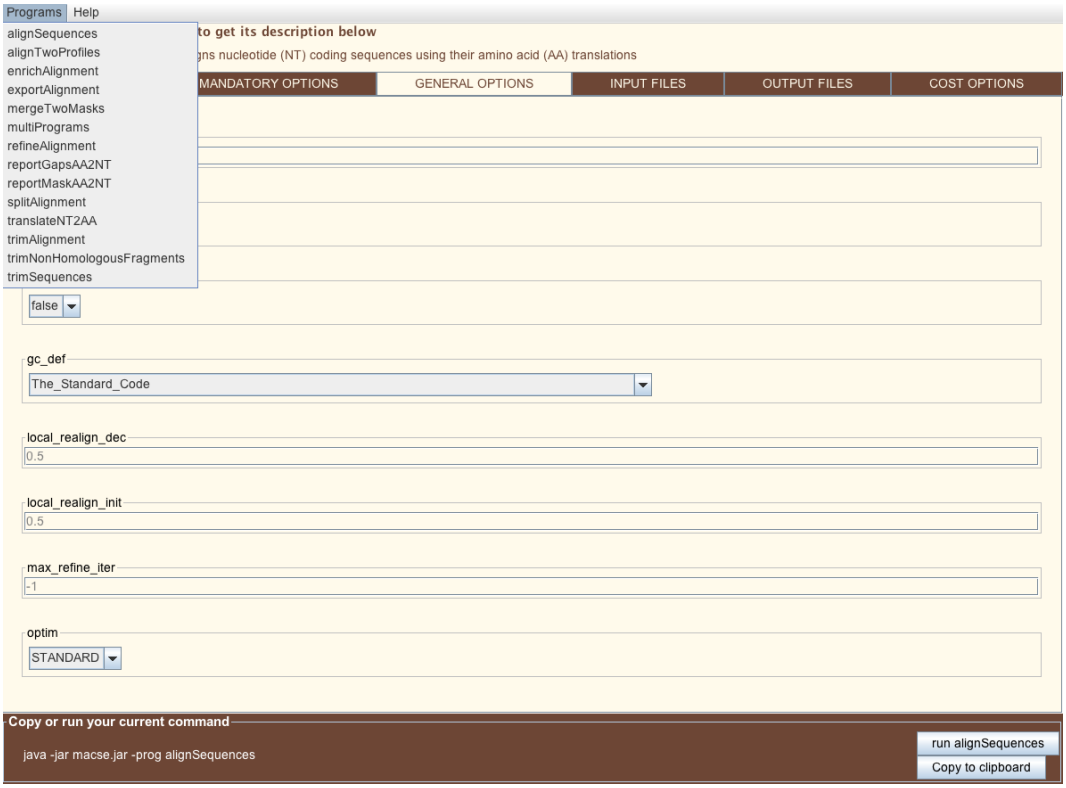


Figure 1 An overview of MACSE v2 Graphical User Interface (GUI) showing the list of available subprograms in the "Programs" menu.

is displayed. Moreover, the corresponding command line appears at the bottom of the GUI and can be directly copied to the clipboard to run the same analysis via the command line and ensure the traceability of the analysis.

The following command launches the command line version of MACSE and prints a help message listing all valid subprograms with a one-line description for each of them:

```
$ java -jar ./macse_v2.03.jar -help
```

The `-prog` option allows users to specify the subprogram to be executed; without any further specification, the following command will print a basic help message for the *alignSequences* subprogram describing its mandatory options:

```
$ java -jar ./macse_v2.03.jar -prog alignSequences
```

Adding the `-help` option will print a help message with more detailed information on the subprogram and the complete list of its available options:

```
$ java -jar ./macse_v2.03.jar -prog alignSequences -help
```

As MACSE is run through the Java virtual machine, the memory that Java is allowed to use can be increased via the `-Xmx` option. This is not a MACSE option *per se*, but it is definitely essential for dealing with relatively large data sets. The following command will for instance allocate 600 MB to the Java virtual machine to run the *alignSequences* subprogram:

```
$ java -jar ./macse_v2.03.jar -Xmx 600m -prog alignSequences
```

2.3:6 Metabarcoding alignements using MACSE

The most basic usage of MACSE to align protein-coding sequences is to use the *alignSequences* subprogram with a `.fasta` file containing protein-coding nucleotide sequences. As *alignSequences* relies on amino acid translations to align sequences, the input nucleotide sequences need to be all in the same forward direction (no reverse complement sequences) and should consist of an open reading frame (ORF) for most of their length (no UTR or intron fragments). The following command will align the 20 mammalian sequences of the *TMEM184a* nuclear gene contained in the `tmem184a.fasta` file (available from the MACSE online tutorial) with default parameters:

```
$ java -jar ./macse_v2.03.jar -prog alignSequences -seq tmem184a.fasta
```

This command will generate two `.fasta` files respectively containing the protein-coding nucleotide sequences aligned as codons and the corresponding amino acid alignment (Figure 2). In this example, MACSE detected frameshifts indicated by exclamation marks ('!') in both the dolphin (*Tursiops*) and the orang-utan (*Pongo*) *TMEM184a* nucleotide sequences. These frameshifts stem from the same two nucleotide deletions and likely correspond to sequencing or annotation errors of this gene in these two species. As MACSE allows aligning these protein-coding sequences by conserving the coding frame, the incomplete codons containing the inferred frameshifts are directly translated into exclamation marks ('!') in the corresponding amino acid alignment.

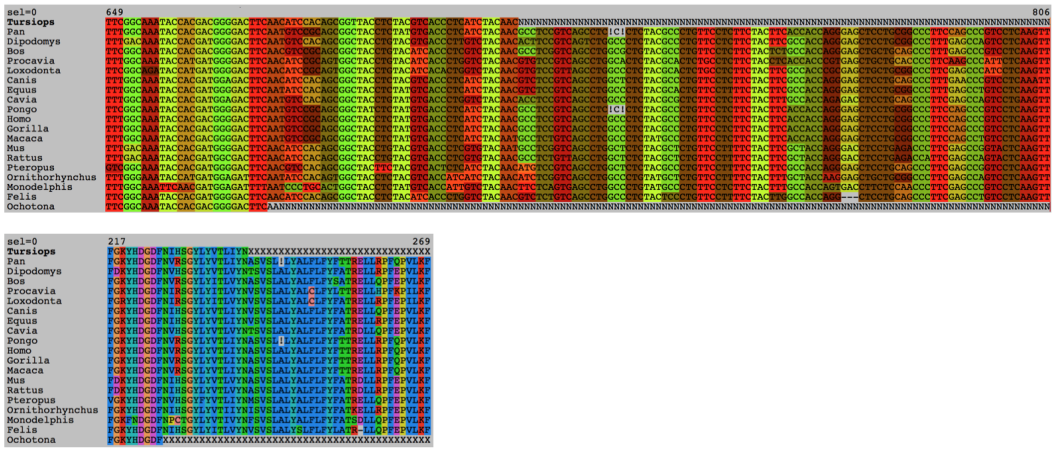


Figure 2 Excerpts of the output nucleotide ("_NT") and amino acid ("_AA") alignments of 20 mammalian sequences of the *TMEM184a* nuclear gene obtained with MACSE subprogram *alignSequences* basic usage and visualized using SeaView (Gouy et al., 2010). Note the frameshifts indicated by exclamation marks ('!') inferred by MACSE in the dolphin (*Tursiops*) and orang-utan (*Pongo*) codon sequences and in the corresponding amino acid alignment.

By default, the names of the output files are based on the input file name by adding the "_NT" and "_AA" suffixes for nucleotides and amino acids, respectively. Custom output file names can be specified using the `-out_NT` and `-out_AA` options. MACSE relies on amino acid translation and uses the standard genetic code by default. Other genetic codes could be specified using the `-gc_def` option following the NCBI nomenclature (`https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi`). For instance, the invertebrate mitochondrial code is number 5. The following command allows aligning mitochondrial *COI* sequences of grasshoppers:

```
$ java -jar ./macse_v2.03.jar
```

```
        -prog alignSequences
        -seq Grasshoppers_COI.fasta
        -gc_def 5
        -out_NT Grasshoppers_COI_NT.fasta
        -out_AA Grasshoppers_COI_AA.fasta
```

If the data set contains sequences with different genetic codes, as it could be the case, for instance, in mitochondrial *COI* barcoding data from multiple animal phyla, they could be specified in a separated text file using the `-gc_file` option. This file should indicate, on each line, the sequence names with their corresponding genetic code numbers. The following command will allow aligning metazoan *COI* sequences extracted from Singh et al. (2009) including five different genetic codes:

```
$ java -jar ./macse_v2.03.jar
          -prog alignSequences
          -seq Singh2009_COI.fasta
          -gc_file Singh2009_COI_gc_file.txt
          -out_NT Singh2009_COI_NT.fasta
          -out_AA Singh2009_COI_AA.fasta
```

A key feature of MACSE for aligning protein-coding barcoding data sets, which are by essence well conserved at the amino acid level, concerns the way "cost parameters" can be fine tuned, in particular the costs associated with frameshifts (`-fs` option) and stop codons (`-stop` option). As in most multiple sequence alignment software, the ratio between gap extension cost (`-gap_ext` option) and gap opening cost (`-gap_op` option) can be specified in MACSE. However, MACSE also allows adjustment of the relative cost of gaps appearing at the sequence extremities (`-gap_op_term` and `-gap_ext_term` options). By default, external gaps are less penalized as they often reflect the fact that a sequence was partially sequenced. Similarly, the occurrence of one or two missing nucleotides at the sequence extremities lead to incomplete codons; but such external frameshifts (`-fs_term` option) should not be as penalized as those occurring in the middle of a sequence. Moreover, when a data set contains a mix of functional and pseudogene sequences, or sequences of variable sequencing qualities (e.g. reference genome sequences and *de novo* assembled contigs), it might be relevant to assign different penalties for the frameshifts (`-fs_lr` option) and stop codons (`-stop_lr` option) appearing in the less reliable sequences. In such cases, MACSE allows the user to *a priori* define two sets of sequences by providing two `.fasta` files as input instead of a single one. The most reliable sequences are in the file provided by the `-seq` option, whereas the least reliable ones are in the file provided by the `-seq_lr` option. The default cost parameters usually work fine, but guidelines to help adjusting parameter costs for some specific types of sequence data sets are provided in the MACSE online documentation. The default values for each parameter can be explored through the GUI.

In the context of metabarcoding studies, one of the main challenges is to add thousands of newly generated barcoding sequences to a reference database for subsequent taxonomic assignment. MACSE v1 was successfully used to implement such an approach in the context of the Moorea Biocode project for characterizing coral reef fish gut contents based on *COI* metabarcoding (Leray et al., 2013). The newly developed *enrichAlignment* subprogram of MACSE v2 can now be used to sequentially enrich a reference alignment with newly generated sequences, possibly adapting the cost parameters for the latter. This subprogram also contains specific options to control the quality of the sequences to be added. For instance, the following command will sequentially enrich a reference alignment of *COI* arthropod sequences by adding only newly generated *COI* fragments obtained from fish gut content that

do not induce too many frameshifts (`-maxFS_inSeq` option), stop codons (`-maxSTOP_inSeq` option), and insertions (`--maxINS_inSeq` option):

```
$ java -jar ./macse_v2.03.jar
          -prog enrichAlignment
          -align Moorea_BIOCODE_small_ref.fasta
          -seq Moorea_BIOCODE_small_ref.fasta
          -seq_lr noctural_diet_sample.fasta
          -gc_def 5
          -fs_lr 10
          -stop_lr 10
          -maxFS_inSeq 0
          -maxINS_inSeq 0
          -maxSTOP_inSeq 1
```

In addition to the two usual output files –respectively containing the final enriched nucleotide ("_NT") and amino acid alignments ("_AA")– the *enrichAlignment* subprogram provides a tabular text file providing detailed statistics for each target sequence, including how many stop codons, frameshifts, and insertion events were required to align it with the reference alignment, and whether it has been added or not based on the specified criteria.

2.2 MACSE_BARCODE: An efficient metabarcoding alignment pipeline

Metabarcoding analysis often requires dealing with several thousands of sequences. Such data sets are not directly tractable with the *alignSequence* subprogram of MACSE, nor by any other classical multiple sequence alignment program (see Chapter 2.2 [Ranwez and Chantret 2020]). However, they can be handled by sequentially adding each barcoding sequence to a reference alignment containing sequences of the targeted locus, as we successfully implemented in the Moorea Biocode project (Leray et al., 2013). Adding sequence one by one is not a second-best option; this strategy has been suggested to produce high quality alignments when dealing with thousands of sequences (Boyce et al., 2014). However, even this strategy could be time consuming when dealing with hundreds of thousands of sequences. Fortunately, at a low taxonomic scale, barcoding sequences are quite similar and contain few indels, if any. As a consequence, if the reference alignment captures most of the sequence diversity, most sequences can be aligned against it without inducing new gap sites in the alignment. This particularity allows the parallelization of the alignment process in which each sequence can be separately aligned to the reference alignment. All sequences that can be aligned to the reference alignment without insertion events can then be combined to build a large alignment containing most sequences of the input data set. This can easily be done using the *enrichAlignment* subprogram of MACSE with the `--maxINS_inSeq 0` option. The remaining sequences can eventually be added afterwards using the same subprogram by relaxing the assumption of no insertion.

The proposed approach implemented in the MACSE_BARCODE pipeline consists of three steps. First, identifying a small subset of a few hundred sequences that best represent the barcoding data set diversity. Second, aligning these representative sequences to build a reference alignment. Third, using this reference alignment to align the thousands of remaining barcode sequences. Online documentation related to MACSE-based pipelines can be found at: https://bioweb.supagro.inra.fr/macse/index.php?menu=docPipeline/docPipelineHtml.

2.2.1 Identifying a small subset of representative sequences

Barcoding data sets may contain non-homologous sequences, as well as reverse complement sequences. To correctly handle these cases, we rely on a carefully chosen reference nucleotide sequence that should translate perfectly into amino acids from start to end. Our pipeline to identify representative sequences (Figure 3) takes advantage of MMseqs2 (Steinegger and Söding, 2017) and takes three key inputs: (1) the fasta file of all barcoding sequences to be aligned, (2) the reference nucleotide sequence, and (3) the genetic code used for translating these sequences. Given these three inputs, the pipeline proceeds as follows:

- Step 1: **MMseqs2 easy-search** is used to compare the translation of each barcode sequence in the six possible reading frames with the translation of the reference nucleotide sequence obtained with the *translateNT2AA* subprogram of MACSE v2. This produces a set of amino acid sequences similar to the reference, and a result table that summarizes all these comparisons in terms of amino acid similarity.
- Step 2: **MMseqs2 easy-cluster** is used to cluster this set of amino acid sequences. The clusters are then sorted based on the number of sequences they contain. The centroid sequences of the largest clusters are identified and their nucleotide sequences are extracted to form the set of representative sequences. The clustering is performed using a strict criterion of 100% amino acid sequence identity. This script has two tuning parameters that control the number of representative sequences. The first one `--in_minClustSize` allows specifying the minimal number of sequences a cluster must contain to be considered (default value = 10). The second one `--in_maxRepresentativeSeqs` allows specifying the maximum number of relevant sequences in the output (default value = 100).
- Step 3: the output of the **MMseqs2 easy-search** step is processed to identify all input sequences that are homologous to the reference sequence and to reverse complement them using seqtk (`https://github.com/lh3/seqtk`), if needed.

The **representative_seqs** sub-pipeline is provided as a Singularity container (Kurtzer et al., 2017) and can be built from the receipt file available from the MACSE_V2_PIPELINES page (`https://github.com/ranwez/MACSE_V2_PIPELINES/tree/master/MACSE_BARCODE`), or directly downloaded from the Singularity official library (Sochat et al., 2017) using the following command:

```
$ singularity pull
        --arch amd64 library://vranwez/default/representative_seqs:v01
```

As for all our Singularity based pipelines (Ranwez et al., 2020), the help message can be obtained by launching them without parameters using the following command:

```
$ ./representative_seqs_v01.sif
```

To launch the identification of the representative sequences for the Magnoliophyta *matK* data set (see Section 3 below) the command line is:

```
$ ./representative_seqs_v01.sif
        --in_refSeq  Magnolia_officinalis_NC_020316.1_matK_ref.fasta
        --in_genetic_code 11
        --in_seqFile Magnoliophyta_BOLD_matK_107413seqs.fasta
        --out_repSeq Magnoliophyta_matK_repSeq.fasta
        --out_homologSeq Magnoliophyta_matK_homologous.fasta
        --out_listRevComp Magnoliophyta_matK_revComp.list
        --in_minClustSize 10
        --in_maxRepresentativeSeqs 100
```

2.3:10 Metabarcoding alignements using MACSE

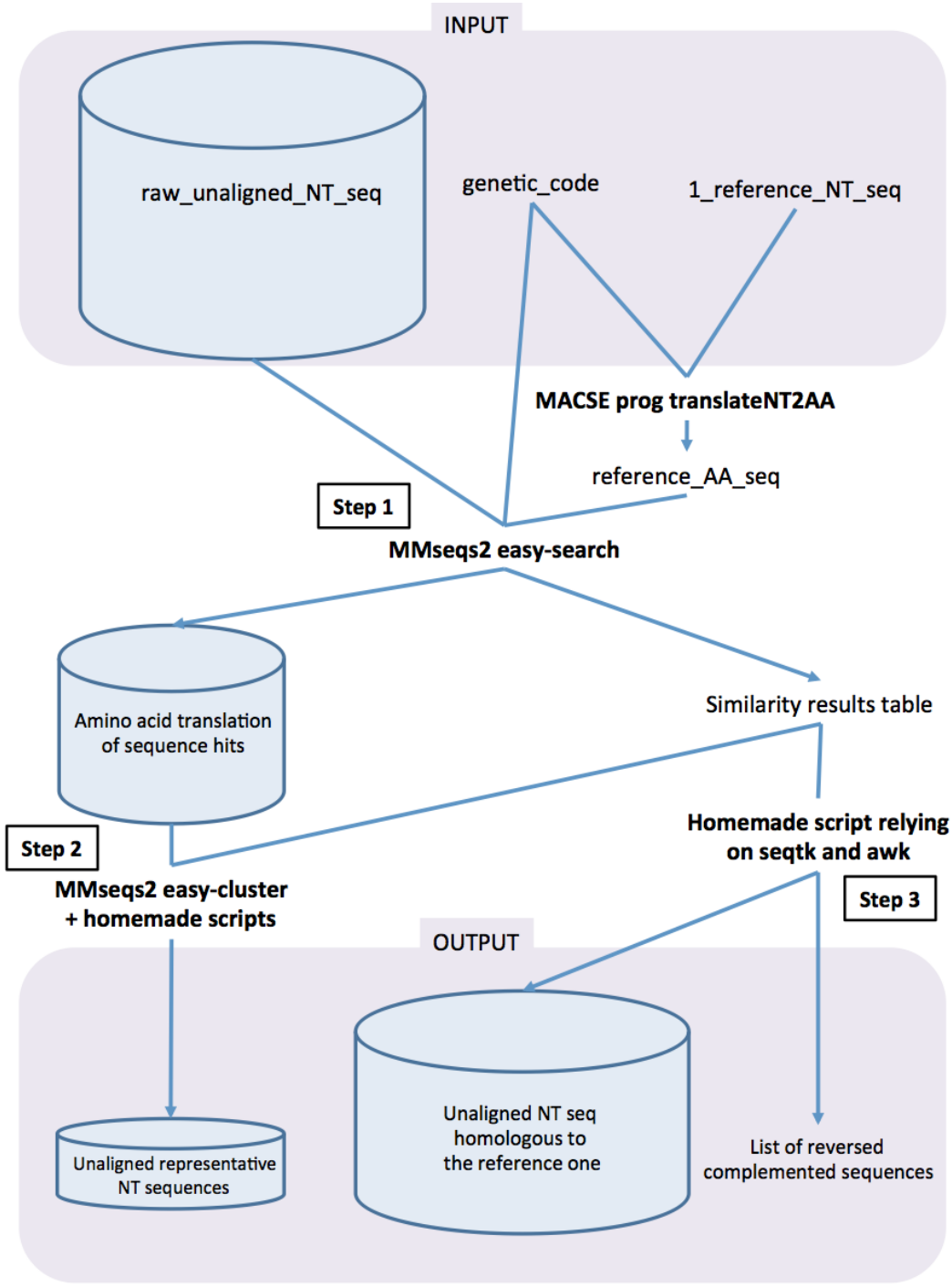


Figure 3 Overview of the representative-sequences identification step of the MACSE_BARCODE pipeline as implemented in the **representative_seqs** Singularity container.

2.2.2 Aligning relevant sequences to get a representative alignment

The relevant nucleotide sequences identified at the previous step could be aligned using the OMM_MACSE pipeline. This pipeline was initially designed to be able to rapidly infer the thousands of CDS alignments of the 10th release of the OrthoMaM database (Scornavacca et al., 2019). This pipeline relies on the amino acid translations to align the coding sequences and includes several optional filtering steps and is more extensively described in Ranwez et al. (2020). It is also provided as a Singularity container and can be built from the receipt file available on our github page or directly downloaded from the Singularity official library using the following command:

```
$ singularity pull
        --arch amd64 library://vranwez/default/omm_macse:v10.02
```

For this specific task, we advise to use the reference nucleotide sequence as the unique "reliable sequence" and to provide the subset of relevant nucleotide sequences as "less reliable sequences". This should help to identify the correct reading frame in case some of the relevant sequences do not start on the first reading frame. We also suggest to avoid the pre-filtering, and alignment trimming steps to ensure that the reference sequence is preserved completely even if the relevant sequences correspond only to a specific sub-fragment:

```
$ ./omm_macse_v10.02.sif
        --in_seq_file Magnolia_officinalis_NC_020316.1_matK_ref.fasta
        --in_seq_lr_file Magnoliophyta_matK_repSeq.fasta
        --out_dir REF_ALIGN_Magnoliophyta_matK
        --out_file_prefix Magnoliophyta_matK
        --genetic_code_number 11
        --no_prefiltering
        --min_percent_NT_at_ends 0
        --java_mem  200m
```

To simplify the process, we provide a Nextflow workflow called **P_buildRefAlignment** that chains these two first steps to directly build the alignment of representative sequences. Nextflow (Di Tommaso et al., 2017) enables scalable and reproducible scientific workflows using software containers allowing the adaptation of pipelines written in the most common scripting languages. It could easily be installed using the following linux commands:

```
$ curl -s https://get.nextflow.io | bash
```

or

```
$ wget -qO- https://get.nextflow.io | bash
```

To launch the **P_buildRefAlignment** workflow, the two previously described Singularity containers **representative_seqs** and **omm_macse** are needed and the "nextflow.config" file should be adapted to the computer environment.

Nextflow separates the workflow itself from the directive regarding the correct way to execute it in the environment. By default, if no "nextflow.config" file is provided, the workflow is launched as a single process on the current machine. One key advantage of Nextflow is that, by changing slightly the "nextflow.config" file, the same workflow will be parallelized and launched to exploit the full resources of a high performance computing (HPC) cluster. The key parameters to change in this configuration file are: (1) the "executor", which could be "local" to run on a standard machine, "sge" or "slurm" to be launched on a HPC cluster or even run on the cloud, and (2) the "queue", which specifies on which queue the job should

be run if a grid based executor is used. An example "nextflow.config" file is provided on the MACSE_BARCODE github page. For instance, the alignment of representative sequences for the Magnoliophyta *matK* data set could be directly built by just running the following command:

```
$ ./nextflow P_buildRefAlignment.nf
        --refSeq Magnolia_officinalis_NC_020316.1_matK_ref.fasta
        --seqToAlign Magnoliophyta_BOLD_matK_107413seqs.fasta
        --geneticCode 11
        --outPrefix Magnoliophyta_matK
```

Note that, despite the care taken in the construction of this pipeline, the output reference alignment may still contain errors. We thus strongly advise to carefully check the resulting alignment, and to remove spurious sequences, if present, before using it as a reference alignment for aligning the remaining thousands of barcode sequences.

2.2.3 Aligning thousands of barcoding sequences using a reference alignment

In order to align the remaining barcode sequences to the reference alignment (Figure 4), a second Nextflow workflow named **P__enrichAlignment** automatizes the two following steps for each sequence. First, each targeted barcoding sequence is compared with the reference alignment and its extremities are trimmed if they are not homologous to the reference alignment. This step allows removing 5' or 3' sequence extremities that do not correspond to the target barcoding locus and, if kept, would impede the sequence to be added without insertion events. Second, the trimmed version of each barcoding sequence is aligned with the reference alignment. The number of unexpected frameshifts, stop codons, and insertion events observed in the aligned sequence is counted and saved in a report `.csv` file, which allows to know exactly the reasons why each sequence has been kept, or not, in the final alignment. By default, the sequences present in the final alignment are those that can be aligned with the reference alignment while having at most two internal frameshifts (incomplete codons at 5' and 3' ends of a sequence are not penalized) and one internal stop codon (a stop codon as the final codon is not penalized). This corresponds to the following options of the *enrichAlignment* subprogram of MACSE `-maxFS_inSeq 2 -maxINS_inSeq 0 -maxSTOP_inSeq 1`. It is easy to spot these parameters in the Nextflow pipeline and to change them if needed. The only parameter that should not be changed is `-maxINS_inSeq 0` as it is a prerequisite for the parallelization. Basically, this workflow (summarized in Figure 4) simply splits the large input data set in small subsets of 100 sequences that are treated in parallel (trimmed then aligned) before concatenating the obtained result files.

To execute this workflow, the previously computed reference alignment, the set of homologous barcode sequences to be aligned, and the genetic code to be used should be provided, as well as a prefix for the output file names using the following command:

```
$ ./nextflow P_enrichAlignment.nf
        --refAlign Magnoliophyta_MATK_reference_alignment_NT.aln
        --seqToAlign Magnoliophyta_MATK_homolog.fasta
        --geneticCode 11
        --outPrefix Magnoliophyta_MATK
```

Finally, if running the **P__buildRefAlignment** and **P__enrichAlignment** workflows separately is advisable to check the identification of the representative sequences and the construction of the reference alignment, we also provide a third Nextflow workflow named

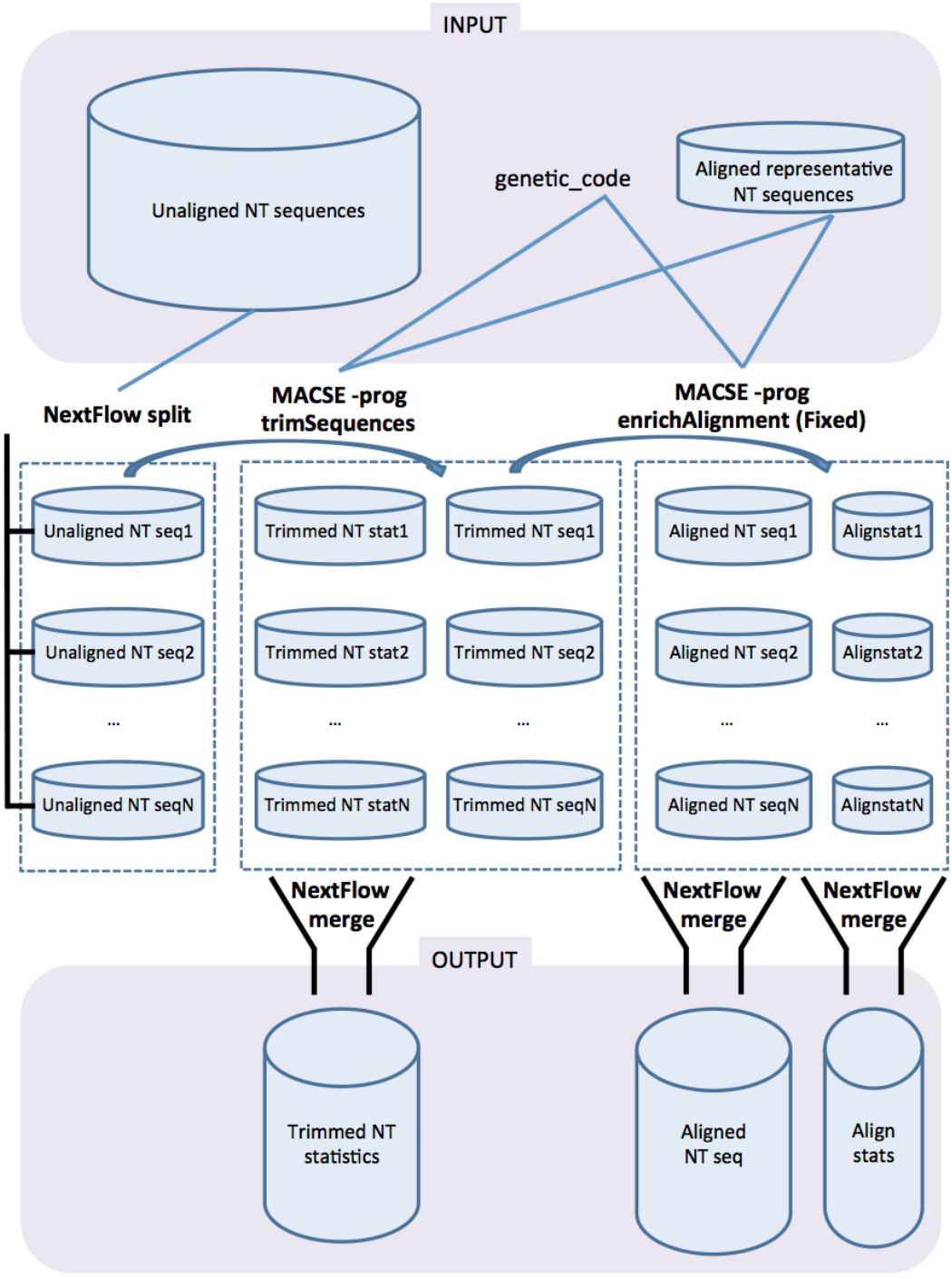


Figure 4 Overview of the parallel enrich alignment step of the MACSE_BARCODE pipeline as implemented in the **P__enrichAlignment** Nextflow workflow.

P__macse__barcode to execute the whole MACSE_BARCODE pipeline. This workflow could be simply executed on the Magnoliophyta *matK* data set by providing the reference sequence, the initial set of barcode sequences to be aligned, the genetic code, and a prefix for

the output file names using the following command:

```
$ ./nextflow P_macse_barcode.nf
        --refSeq Magnolia_officinalis_NC_020316.1_matK_ref.fasta
        --seqToAlign Magnoliophyta_BOLD_matK_107413seqs.fasta
        --geneticCode 11
        --outPrefix Magnoliophyta_matK
```

3 (Meta)barcoding case studies

3.1 MACSE automatically detects sequencing errors in *COI* sequences

3.1.1 Bad quality crayfish *COI* sequences

The first example data set (TK2006) is an alignment of 89 crayfish *COI* sequences produced by Taylor and Knouft (2006). This data set includes 24 poor quality sequences that have been flagged as "COI-like" by GenBank after submission as it has been pointed out by Buhay (2009), who later provided a manually curated alignment for these sequences. We used this data set to evaluate the capability of MACSE to automatically provide an accurate alignment of a set of sequences containing both correct and incorrect *COI* sequences. As was done manually by Buhay (2009), we used the *COI* sequence extracted from the complete mitochondrial genome of *Cherax destructor* (NC_011243) as a reference to automatically align the TK2006 data set using the *alignSequences* subprogram of MACSE. We used differential costs for frameshifts (`-fs 30`) and stop codons (`-stop 30`) for the reference sequence, and `-fs_lr 10` and `-stop_lr 10` for all other sequences considered as less reliable including "COI-like" sequences:

```
$ java -jar ./macse_v2.03.jar
        -prog alignSequences
        -gc_def 5
        -seq Cherax_destructor_ref.fasta
        -seq_lr TK2006.fasta
        -fs 30
        -stop 30
        -fs_lr 10
        -stop_lr 10
        -out_NT TK2006_macse_NT.fasta
        -out_AA TK2006_macse_AA.fasta
```

The resulting nucleotide and amino acid alignments can be visually inspected using SeaView, which handles the exclamation mark ('!') character used by MACSE to identify frameshifts and also allows coloring the alignment by codons (Figure 5).

The statistics on the number of frameshifts and stop codons inferred in each sequence of the resulting nucleotide alignment can be output in a `.csv` file (`--out_stat_per_seq` option) using the *exportAlignment* subprogram with the following command:

```
$ java -jar ./macse_v2.03.jar
        -prog exportAlignment
        -gc_def 5
        -align TK2006_macse_NT.fasta
        -out_stat_per_seq TK2006_macse_NT_stat.csv
```

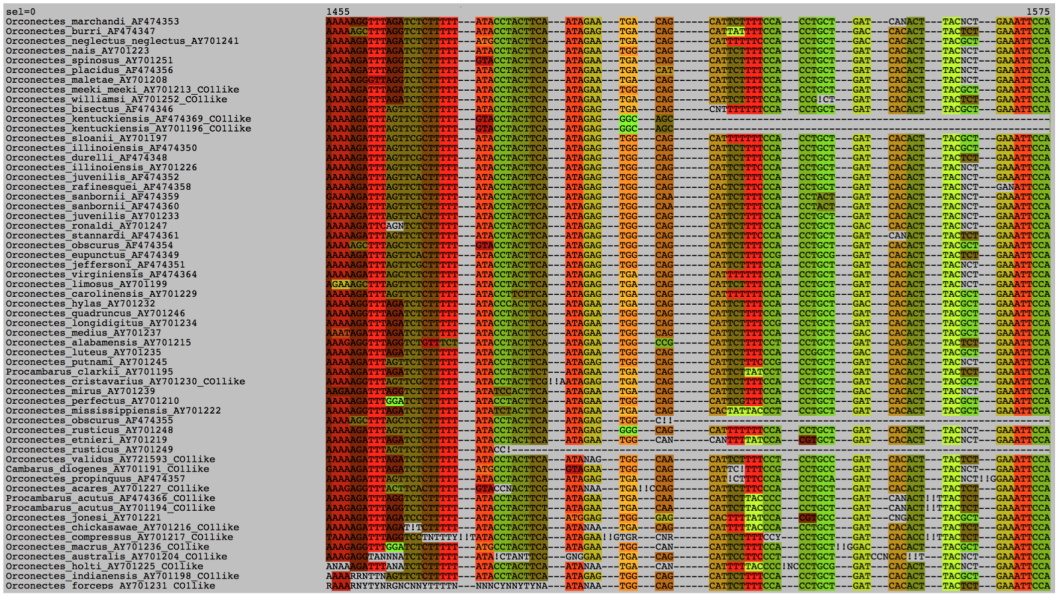


Figure 5 Excerpt of the crayfish *COI* nucleotide data set of Taylor and Knouft (2006) aligned by MACSE and visualized by SeaView using codon colors. Note the numerous frameshifts ('!') inferred by MACSE in the bad quality "COI-like" sequences.

MACSE inferred 20 sequences containing frameshifts including 18 annotated as "COI-like", but also two other sequences, *Orconectes margorectus* (AF474362) and *O. propinquus* (AF474357), containing frameshifts caused by additional nucleotides. Frameshifts in those sequences have probably gone unnoticed because they both occur close to the end of the sequences and do not lead to stop codons in the few nucleotides following the insertions. MACSE allows automatically spotting such cryptic cases. The worst "COI-like" sequence was *O. australis* (AY701204) that contains four frameshifts and an internal stop codon due to multiple missing and additional nucleotides. For six "COI-like" sequences, MACSE did not infer the presence of any frameshifts or stop codons: *O. inermis* (AY701201), *O. kentuckiensis* (AF474369 and AY701196), *O. meeki meeki* (AY701213), *O. neglectus chaenodactylus* (AY701240) and *O. pellucidus* (AY701203). Careful visual inspection showed that most of these problems are likely stemming from sequencing errors rather than representing signs of pseudogenization (*numts*).

3.1.2 "COI-like" crustacean sequences in GenBank

The second example data set (BT2009) corresponds to the crustacean "COI-like" sequences harvested from GenBank and presented in Table 1 of Buhay (2009). This data set contains a mix of 54 "COI-like" complete and partial sequences. These sequences were manually assessed in great detail by Buhay (2009) including three sequences that were determined to likely not represent genuine *COI* sequences. We used this example data set to evaluate the capacity of MACSE to automatically detect sequencing errors that have been previously assessed by an expert eye. The *trimNonHomologousFragments* subprogram of MACSE allows identifying and trimming non homologous sequence fragments before further alignment. It could be used to identify entirely non-homologous sequences since such sequences will be trimmed along their entire lengths. A visualization of the result of this pre-filtering process can be output

in a `.fasta` file (`-out_mask_detail` option) in which the removed nucleotides are written in lowercase letters. The *trimNonHomologousFragments* subprogram also outputs a `.csv` file detailing the statistics of this pre-filtering process on each sequence (`-out_trim_info` option). This file contains the number of nucleotides and the proportion of each sequence that have been removed:

```
$ java -jar ./macse_v2.03.jar
       -prog trimNonHomologousFragments
       -gc_def 5
       -seq BT2009.fasta
       -out_mask_detail BT2009_trim_mask.fasta
       -out_trim_info BT2009_trim_stats.csv
       -out_NT BT2009_trim_NT.fasta
       -out_AA BT2009_trim_AA.fasta
```

This non-homologous fragment-trimming step allowed to automatically detect the three sequences that are likely not true mitochondrial *COI* sequences: *Chthamalus dalli* (AY795367), *Farfantepenaeus subtilis* (AY344198), and *Fenneropenaeus indicus* (AY395245). Indeed, MACSE automatically removed these sequences in their entirety. In addition, six other sequences were found to contain divergent fragments representing up to 30% of their length that were masked in the resulting output: *Scopelocheirus schellenbergi* (AY830432), *Parastacoides tasmanicus* (AF482492), *Orconectes indianensis* (AY701198), *Orconectes forceps* (AY701231), *Liberonautes chaperi* (AF399977), and *Caligus* sp. (EF452643). Most of these sequences were spotted as very low quality sequences by Buhay (2009) with sloppy 5' or 3' ends and lots of ambiguities.

In order to further assess the quality of the remaining 51 "COI-like" sequences, we aligned the masked sequences against the same reference *COI* sequence of *Cherax destructor* (NC_011243) previously used for the TK2006 data set. We used differential costs for frameshifts (`-fs 30`) and stop codons (`-stop 30`) for the reference sequence, and `-fs_lr 10` and `-stop_lr 10` for the less reliable "COI-like" sequences:

```
$ java -jar ./macse_v2.03.jar
       -prog alignSequences
       -gc_def 5
       -seq Cherax_destructor_ref.fasta
       -seq_lr BT2009_trim_NT.fasta
       -fs 30
       -stop 30
       -fs_lr 10
       -stop_lr 10
       -out_NT BT2009_trim_macse_NT.fasta
       -out_AA BT2009_trim_macse_AA.fasta
```

The resulting nucleotide alignment was visually inspected using SeaView (Figure 6).

The statistics on the number of frameshifts and stop codons inferred per sequence in these alignments were computed using the *exportAlignment* subprogram of MACSE:

```
$ java -jar ./macse_v2.03.jar
       -prog exportAlignment
       -gc_def 5
       -align BT2009_trim_macse_NT.fasta
       -out_stat_per_seq BT2009_trim_macse_NT_stat.csv
```

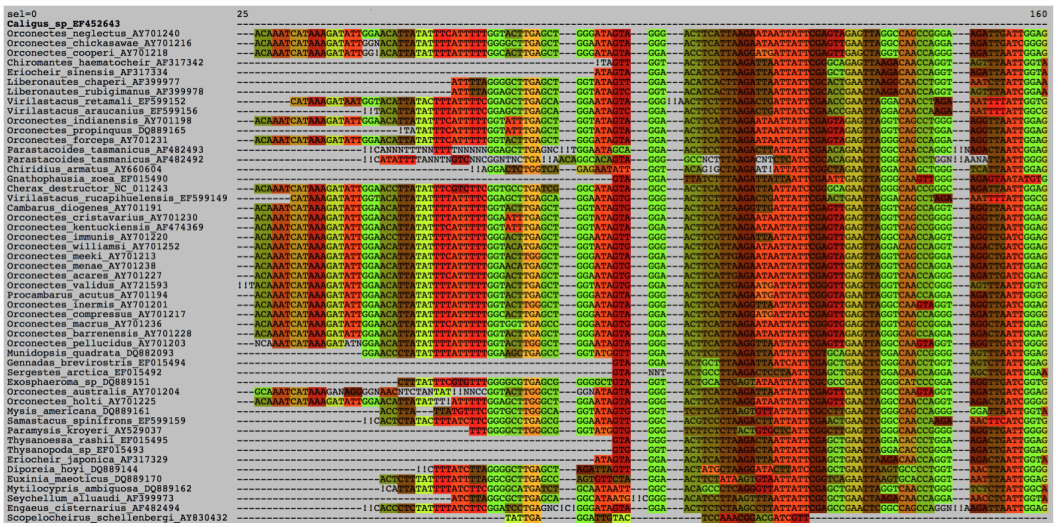


Figure 6 Excerpt of the crustacean "COI-like" data set assembled from GenBank by Buhay (2009) trimmed and aligned by MACSE, and visualized by SeaView using codon colors. Note the numerous frameshifts ('!') inferred by MACSE in the bad quality sequences.

In all cases but three, MACSE correctly retrieved the problematic frameshifts and stop codons previously identified manually by Buhay (2009). The three exceptions concern sequences containing frameshifts occurring very close from sequence extremities (*Orconectes inermis* AY701201 and *Orconectes macrus* AY701236) and an internal frameshift caused by a 1-bp insertion compensated by a 1-bp deletion few nucleotides downstream (*Orconectes meeki* AY701213).

3.2 MACSE alleviates the impact of *numts* in COI-based barcoding species delimitation

Grasshoppers and crayfish are two invertebrate groups that are well known for presenting a high rate of nuclear integration of mitochondrial DNA (mtDNA) that gives rise to nuclear mitochondrial pseudogenes (*numts*). These *numts* are frequently co-amplified with the PCR targeted mtDNA sequences and could strongly impact barcoding analyses based on the mitochondrial *COI* marker. Song et al. (2008) studied the effect of *numts* on COI-based phylogenetic and barcoding analyses. They used PCR with universal arthropods barcoding primers to amplify *COI* in grasshoppers and crayfish. They showed that universal barcode PCR primers co-amplified numerous *COI numts* in both groups resulting in large overestimates of the number of species that can be delineated by *COI* barcoding. Their grasshopper data set (SG2008) contains 95 *COI* sequences including 88 co-amplified *numts* and their crayfish data set (SC2008) comprises 183 *COI* sequences with 101 co-amplified *numts*. *Numts* accumulate frameshifts and stop codons because they become nonfunctional after nuclear integration and are no longer under selective pressure to conserve the open reading frame. We used these two data sets to illustrate the efficiency of MACSE to identify *numts* based on the presence of frameshifts and stop codons as an automatic quality control step aimed at removing these pseudogene sequences before conducting downstream species delimitation analyses. The SG2008 and SC2008 data sets were aligned using the *alignSequences* subprogram of MACSE without *a priori* specification of potentially less

reliable *numts* sequences and with default frameshift (`-fs 30`) and stop codon (`-stop 30`) costs for all sequences:

Grasshoppers:

```
$ java -jar ./macse_v2.03.jar
          -prog alignSequences
          -gc_def 5
          -seq SG2008.fasta
          -fs 30
          -stop 30
          -out_NT SG2008_macse_NT.fasta
          -out_AA SG2008_macse_AA.fasta
```

Crayfish:

```
$ java -jar ./macse_v2.03.jar
          -prog alignSequences
          -gc_def 5
          -seq SC2008.fasta
          -fs 30
          -stop 30
          -out_NT SC2008_macse_NT.fasta
          -out_AA SC2008_macse_AA.fasta
```

The statistics on the number of frameshifts and stop codons inferred for each sequence in the resulting alignments were computed using the *exportAlignment* subprogram of MACSE:

Grasshoppers:

```
$ java -jar ./macse_v2.03.jar
          -prog exportAlignment
          -gc_def 5
          -align SG2008_macse_NT.fasta
          -out_stat_per_seq SG2008_macse_NT_stat.csv
```

Crayfish:

```
$ java -jar ./macse_v2.03.jar
          -prog exportAlignment
          -gc_def 5
          -align SC2008_macse_NT.fasta
          -out_stat_per_seq SC2008_macse_NT_stat.csv
```

For the grasshoppers data set, MACSE detected 37 out of the 95 sequences containing at least one frameshift and/or stop codon, and 99 out of the 183 sequences for the crayfish data set. The potential *numt* sequences containing at least one frameshift and/or one stop codon were then removed from the resulting nucleotide alignments. Ambiguously aligned and highly incomplete codon sites were excluded using Gblocks (Castresana, 2000) with default relaxed codon parameters for both the complete and reduced MACSE alignments. Maximum likelihood (ML) phylogenetic inference was then conducted on the four resulting data sets with PhyML v3.1 (Guindon et al., 2009) using a GTR+G8 model and SPR branch swapping on a BIONJ starting tree. The ML phylograms obtained for the grasshoppers and crayfish data sets are presented in Figures 7 and 8, respectively.

In order to gauge the impact of automatically removing *numts* using MACSE on species delimitation, we applied the multi-rate Poisson Tree Process model (Kapli et al. 2017) on the four resulting ML phylograms using the mPTP server (`http://mptp.h-its.org`).

Figure 7 Maximum likelihood phylograms obtained before (A) and after (B) filtering the grasshoppers data set of Song et al. (2008) using MACSE to detect *numt* sequences containing frameshits and/or stop codons. In this case, MACSE automatically removed 37 *numts*.

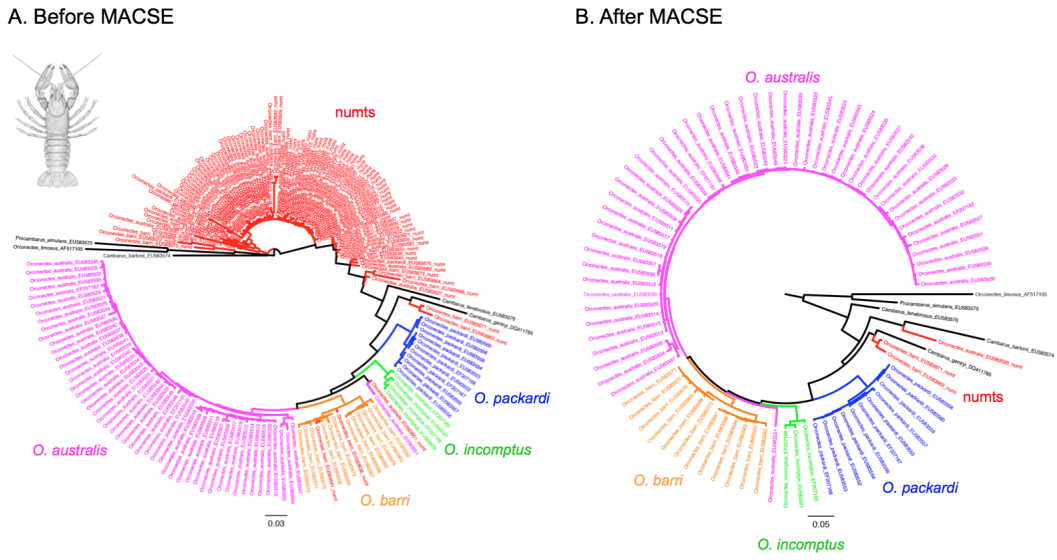


Figure 8 Maximum likelihood phylograms obtained before (A) and after (B) filtering the crayfish data set of Song et al. (2008) using MACSE to detect *numt* sequences containing frameshits and/or stop codons. In this case, MACSE automatically removed 99 *numts* (in red).

For grasshoppers, mPTP delimited 20 species from the 95 sequences of the original data set, which included 88 *numts*. The number of delimited species was halved to 10 with the 58 sequences retained by MACSE that still included 51 potential *numts*, which did not contain any frameshift or stop codon. For crayfish, mPTP delimited 26 species from the 183 sequences of the original data set, which included 101 *numts*. Again, the number of delimited species dropped to 8 with the 84 sequences retained by MACSE that only included 3 potential *numts*, which did not contain any frameshift or stop codon.

3.3 MACSE_BARCODE accurately aligns thousands of metabarcoding sequences

In order to illustrate the efficiency of the MACSE_BARCODE pipeline, we applied the two Nextflow workflows **P_buildRefAlignment** and **P_enrichAlignment** to barcode sequences publicly available for ants, mammals, and flowering plants, which were downloaded through the Taxonomy portal of the BOLD database v4 (http://v4.boldsystems.org/index.php/TaxBrowser_Home) on March 3rd 2020 (Table 1).

	BOLD sequences			Homologous sequences		
	per taxa	per taxa and marker	homologous to reference (reverse complemented)	in final alignment	with internal frameshifts	with internal stop codons
Mammalia *COI*	141,145	121,180	117,547 (6)	117,363	223	82
Formicidae *COI*	124,067	121,954	121,792 (33)	121,494	557	16
Magnoliophyta *rbcL*	339,948	121,989	121,598 (116)	121,302	825	346
Magnoliophyta *matK*	339,948	107,413	107,032 (614)	63,250	1,824	143

Table 1 Descriptive statistics of the four BOLD barcoding data sets on which the MACSE_BARCODE pipeline has been applied to construct reference alignments.

3.3.1 Mammalian *COI* sequences in BOLD

As a first example, we aimed at constructing a reference alignment of *COI* barcode sequences for all mammals represented in the BOLD database. As mammalian *COI* sequences are well conserved at the scale of mammals, this first data set serves as an ideal first test case for our approach. Using the Taxonomy portal of the BOLD system v4 (http://v4.boldsystems.org/index.php/TaxBrowser_Home), we downloaded the 141,145 publicly available sequences in the Mammalia section. These raw sequences contain sequences from different molecular markers and also include gaps. Sequences corresponding to *COI* can thus be counted using the following command:

```
$ grep -c COI Mammalia_BOLD_141145seq_raw.fasta
```

This resulted in 121,180 *COI* sequences that were extracted and stored in a new fasta file using the following command:

```
$ grep -A1 COI Mammalia_BOLD_141145seq_raw.fasta
   > Mammalia_BOLD_121180seq_COI_raw.fasta
```

At this stage, gaps could be removed from all sequences and illegal characters such as pipes '|' and spaces could be replaced with underscores "_" to ease further bioinformatic processing using for instance:

```
$ sed -e '/>/!s/-//g'
      -e '/>/s/[| :().,;#]/_/g' Mammalia_BOLD_121180seq_COI_raw.fasta
      > Mammalia_BOLD_121180seq_COI.fasta
```

Using the *Homo sapiens* (NC_012920) full length *COI* sequence as a reference sequence, the alignment of representative sequences for the Mammalia *COI* data set could be built using the **P_buildRefAlignment** workflow by running the following command:

```
$ ./nextflow P_buildRefAlignment.nf
        --refSeq Homo_sapiens_NC_012920_COI_ref.fasta
        --seqToAlign Mammalia_BOLD_121180seq_COI.fasta
        --geneticCode 2
        --outPrefix Mammalia_COI
```

This generates a result folder `RESULTS_REFA_Mammalia_COI` containing a `.fasta` file of unaligned representative sequences (`Mammalia_COI_repSeq.fasta`) and the corresponding nucleotide (`Mammalia_COI_final_align_NT.aln`) and amino acid (`Mammalia_COI_final_align_AA.aln`) alignments. A `.fasta` file containing the barcode sequences identified to be homologous to the reference (`Mammalia_COI_homolog.fasta`) and a list of the names of the sequences that have been reverse complemented (`Mammalia_COI_RevComSeqId.list`) are also provided. In this case, 117,547 sequences (97.0%) were found homologous to the *COI* reference sequence and only six sequences had to be reverse complemented to be aligned (Table 1).

The final alignment of all homologous *COI* sequences could then be computed using the **P_enrichAlignment** workflow by providing the previously computed alignment of representative sequences as a reference alignment (`Mammalia_COI_reference_alignment_NT.aln`) and the set of homologous barcode sequences remaining to be aligned (`Mammalia_COI_homolog.fasta`), and executing the following command:

```
$ ./nextflow P_enrichAlignment.nf
        --refAlign Mammalia_COI_final_align_NT.aln
        --seqToAlign Mammalia_COI_homolog.fasta
        --geneticCode 2
        --outPrefix Mammalia_COI
```

This produces a result folder `RESULTS_ENRICH_Mammalia_COI` containing the nucleotide (`Mammalia_COI_alignAll_NT.aln`) and amino acid (`Mammalia_COI_alignAll_AA.aln`) final alignments of all *COI* sequences, as well as versions of those alignments in which frameshifts ('!') have been removed (`Mammalia_COI_alignAll_NT_exp_noFS.aln`) and (`Mammalia_COI_alignAll_AA_exp_noFS.aln`). Moreover, statistics on the steps of sequence trimming (`Mammalia_COI_preTrimingStat.csv`) and alignment enrichment (`Mammalia_COI_enrich_info.csv`) are provided as `.csv` files so that the fate of each initial sequence can be monitored. Here, the final mammal *COI* alignment (Figure 9) contains 117,363 of the initial 121,180 sequences (96.9%). The alignment of 223 of these sequences required the inference of an internal frameshift and 82 sequences have been integrated in this final alignment while having an internal stop codon (Table 1). These sequences could easily have been excluded from the alignment, if needed.

These analyses have been run on a HPC cluster. According to Nextflow, the identification of the representative sequences and the generation of the reference alignment with **P_buildRefAlignment** took only 8 minutes. The obtention of the final alignment with **P_enrichAlignment** required about 134 hours of CPU time but the final result was produced in just 1 hour and 38 minutes thanks to the parallelization used in the pipeline.

2.3:22 Metabarcoding alignements using MACSE

The whole MACSE_BARCODE pipeline could also be executed directly using:

```
$ ./nextflow P_macse_barcode.nf
        --refSeq Homo_sapiens_NC_012920_COI_ref.fasta
        --seqToAlign Mammalia_BOLD_121180seq_COI.fasta
        --geneticCode 2
        --outPrefix Mammalia_COI
```

Figure 9 Excerpt of the final Mammalia *COI* nucleotide alignment containing 117,363 sequences produced by MACSE_BARCODE including both full-length (1,548 bp) and shorter *COI* fragments as visualized by AliView (Larsson, 2014).

3.3.2 Ant *COI* sequences in BOLD

As a second example, we considered *COI* barcode sequences from all ant specimens represented in the BOLD database. A total of 124,067 publicly available sequences were downloaded from the Formicidae section using the Taxonomy portal of the BOLD system v4. As for mammals, these raw sequences contain sequences from different molecular markers. So, sequences corresponding to *COI* have to be counted:

```
$ grep -c  COI Formicidae_BOLD_124067seq_raw.fasta
```

The resulting 121,954 *COI* sequences were then extracted and stored in a new `.fasta` file:

```
$ grep -A1  COI Formicidae_BOLD_124067seq_raw.fasta
   > Formicidae_BOLD_121954seq_COI_raw.fasta
```

After gap removal and name cleaning, the alignment of representative sequences for the Formicidae *COI* data set was built with the **P__buildRefAlignment** workflow using the full length *COI* sequence of *Solenopsis geminata* (NC_014669.1) as a reference:

```
$ ./nextflow P_buildRefAlignment.nf
        --refSeq Solenopsis_geminata_NC_014669_COI_ref.fasta
        --seqToAlign Formicidae_BOLD_121954seq_COI.fasta
        --geneticCode 5
        --outPrefix Formicidae_COI
```

In this ant *COI* data set, 121,792 sequences were considered to be homologous to the *COI* reference sequence, and 33 sequences had to be reverse complemented to be aligned (Table 1).

The final alignment of all homologous *COI* sequences was then computed using the **P__enrichAlignment** workflow:

```
$ ./nextflow P_enrichAlignment.nf
        --refAlign Formicidae_COI_final_align_NT.aln
        --seqToAlign Formicidae_COI_homolog.fasta
        --geneticCode 5
        --outPrefix Formicidae_COI
```

The final ant *COI* alignment (Figure 10) comprises 121,494 of the initial 121,954 sequences (99.6%). The alignment of 557 of these sequences required the inference of an internal frameshift and 16 sequences were integrated in this alignment while presenting an internal stop codon (Table 1).

3.3.3 Flowering plant *rbcL* sequences in BOLD

In this third example, we considered another taxonomic group and barcoding marker with the chloroplastic *rbcL* gene, which is the first official barcoding marker for flowering plants. The Taxonomy section of the BOLD system v4 public database contains 339,948 raw public sequences for Magnoliophyta. This included *rbcL* sequences but also other barcoding markers such as the chloroplastic *matK* gene. After counting *rbcL* sequences:

```
$ grep -c  rbcL Magnoliophyta_BOLD_339948seq_raw.fasta
```

The resulting 121,989 *rbcL* sequences were extracted and stored in a new `.fasta` file:

```
$ grep -A1 rbcL Magnoliophyta_BOLD_339948seq_raw.fasta
   > Magnoliophyta_BOLD_121989seq_rbcL_raw.fasta
```

After gap removal and name cleaning, the reference alignment of representative sequences for the Magnoliophyta *rbcL* data set was built with the **P__buildRefAlignment** workflow using the *Magnolia officinalis* (NC_020316.1) full length *rbcL* sequence as a reference:

```
$ ./nextflow P_buildRefAlignment.nf
        --refSeq Magnolia_officinalis_NC_020316.1_rbcL_ref.fasta
        --seqToAlign Magnoliophyta_BOLD_121989seq_rbcL.fasta
        --geneticCode 11
        --outPrefix Magnoliophyta_rbcL
```

In this flowering plant *rbcL* data set, 121,598 sequences were found homologous to the *rbcL* reference sequence among which 116 sequences had to be reverse complemented to be aligned (Table 1). The final alignment of all homologous *rbcL* sequences was then computed using the **P__enrichAlignment** workflow:

2.3:24 Metabarcoding alignements using MACSE

Figure 10 Excerpt of the final Formicidae *COI* metabarcoding nucleotide alignment containing 121,494 sequences produced by MACSE_BARCODE including both full-length (1,533 bp) and shorter *COI* fragments as visualized by AliView.

```
$ ./nextflow P_enrichAlignment.nf
        --refAlign Magnoliophyta_rbcL_final_align_NT.aln
        --seqToAlign Magnoliophyta_rbcL_homolog.fasta
        --geneticCode 11
        --outPrefix Magnoliophyta_rbcL
```

The final flowering plant *rbcL* alignment (Figure 11) comprises 121,302 of the initial 121,989 sequences (99.4%). The alignment of 857 of these sequences required the inference of an internal frameshift and 346 sequences were integrated in this alignment despite the presence of an internal stop codon (Table 1). While this *rbcL* alignment has almost the same number of sequences (121,302) as the alignments obtained for mammals (117,363) and ants (121,494) using *COI*, it contains much more sequences comporting an internal frameshift (825 *versus* 223 and 557, respectively) or an internal stop codon (346 *versus* 82 and 16, respectively). This likely indicates that the sequences available for *rbcL* are of lower quality than those available for *COI*. The high number of sequences containing a stop codon is especially surprising as this should be something easy to check before including a sequence in the BOLD database. For instance, the amino acid sequence displayed on BOLD in the detailed record for the sequence GBVC3450-11, which is flagged as mined from GenBank, includes a stop codon without explicit warning (http://www.boldsystems.org/index.php/Public_RecordView?processid=GBVC3450-11).

Figure 11 Excerpt of the final Magnoliophyta *rbcL* metabarcoding nucleotide data set containing 121,302 sequences produced by MACSE_BARCODE including both full-length (1,440 bp) and shorter *rbcL* fragments as visualized by AliView.

3.3.4 Flowering plant *matK* sequences in BOLD

For this last example, we considered the second official barcoding marker for flowering plants with the chloroplastic *matK* gene. The *matK* sequences were counted from the previously downloaded raw Magnoliophyta sequences from BOLD:

```
$ grep -c matK Magnoliophyta_BOLD_339948seq_raw.fasta
```

The resulting 107,413 *matK* sequences were extracted and stored in a new `.fasta` file:

```
$ grep -A1 matK Magnoliophyta_BOLD_339948seq_raw.fasta
   > Magnoliophyta_BOLD_107413seq_matk_raw.fasta
```

After gap removal and name cleaning, the reference alignment of representative sequences for the Magnoliophyta *matK* data set was built with the **P_buildRefAlignment** workflow using the *Magnolia officinalis* (NC_020316.1) full length *matK* sequence as a reference:

```
$ ./nextflow P_buildRefAlignment.nf
    --refSeq Magnolia_officinalis_NC_020316.1_matK_ref.fasta
    --seqToAlign Magnoliophyta_BOLD_107413seq_matK.fasta
    --geneticCode 11
    --outPrefix Magnoliophyta_matK
```

In this flowering plant *matK* data set, 107,032 sequences were found homologous to the reference sequence, 614 of which had to be reverse complemented to be aligned (Table 1).

The proportion of sequences provided in the wrong orientation (0.5%) was much higher than for the other data sets (e.g. 0.005% for the mammal *COI* data set).

The final alignment of all homologous *matK* sequences was then computed using the **P_enrichAlignment** workflow:

```
$ ./nextflow P_enrichAlignment.nf
    --refAlign Magnoliophyta_matK_final_align_NT.aln
    --seqToAlign Magnoliophyta_matK_homolog.fasta
    --geneticCode 11
    --outPrefix Magnoliophyta_matK
```

The final flowering plant *matK* alignment (Figure 12) comprises only 63,250 of the initial 107,413 sequences (58.9%). The alignment of 1,824 of these sequences required the inference of an internal frameshift and 143 sequences were aligned while presenting an internal stop codon (Table 1).

Figure 12 Excerpt of the final Magnoliophyta *matK* metabarcoding nucleotide alignment containing 63,250 sequences produced by MACSE_BARCODE including both full-length (1,536 bp) and shorter *matK* fragments as visualized by AliView.

The fact that more than 40% of the initial *matK* sequences were excluded from the final alignment could reflect the limit of our approach, the poor quality of the sequences available in BOLD for this marker, or more likely a mix of both causes. In fact, many sequences were not inserted because their presence would induce many insertions relative to the reference alignment. This illustrates one limit of our current approach that is based on the conservation of the reference sequence in terms of both amino acid divergence and indel occurrence. Indeed, *matK* is much more variable than *rbcL*, notably in terms of indels (CBOL Plant Working

Group, 2009). However, this could be alleviated by applying the MACSE_BARCODE pipeline at lower taxonomic levels such as the Family level at which *matK* sequences might be more conserved in length. Meanwhile, it seems that the sequences of this data set are of lower quality comparable to the other three data sets with a much higher proportion of the aligned sequences requiring the inference of at least one internal frameshift to be correctly aligned. To ensure that this was not an error of our pipeline, we extracted from the final *matK* alignment the 1,940 sequences (out of a total of 63,250) that were included despite presenting an internal frameshift or a stop codon. This allowed to confirm that, in most cases, the presence of the inferred frameshifts was accurate and that some stop codons appeared right in the middle of the sequences (Figure 13). Altogether, these observations indicate that numerous flowering plant *matK* sequences in BOLD seem to be of relatively poor quality or might represent interesting cases of biologically relevant shifts in translation reading frame, as recently uncovered in Orchidaceae (Barthet et al., 2015).

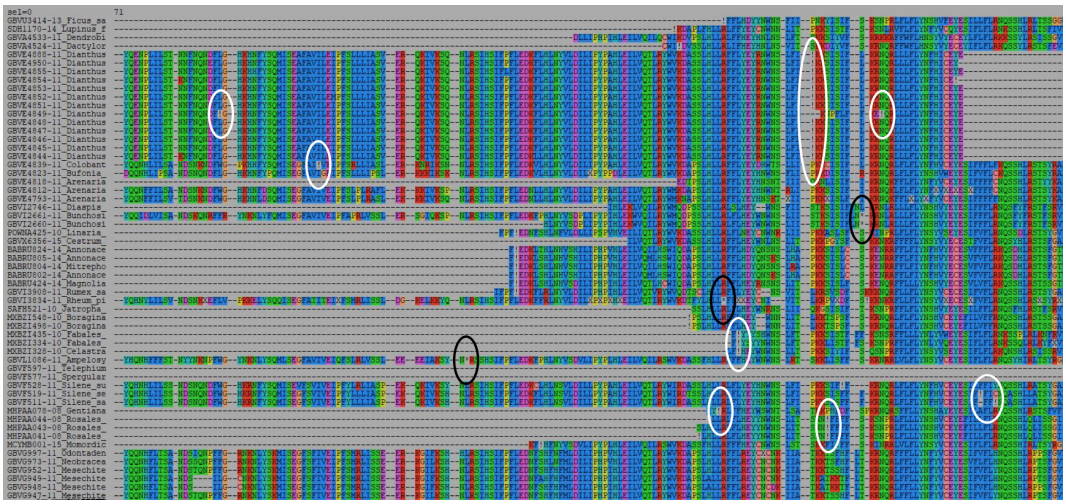


Figure 13 Excerpt of the final Magnoliophyta *matK* final amino acid alignment containing 63,250 sequences focusing on sequences presenting internal frameshifts (white ellipses) and stop codons (black ellipses) as visualized by SeaView.

4 Conclusion

The reference barcoding alignments produced with the MACSE_BARCODE pipeline can be downloaded from the MACSE webpage (`https://bioweb.supagro.inra.fr/macse/index.php?menu=downloadTuto`). Future reference alignments for additional taxonomic groups available in the BOLD database will be distributed through the same webpage. The availability of these quality-controlled alignments for the main protein-coding barcode genes should leverage the power of phylogenetics for taxonomic assignment by allowing to implement probabilistic evolutionary placement in the ever growing range of metabarcoding applications.

Acknowledgements

The project was granted access to the INRA MIGALE bioinformatics platform (`https://migale.inra.fr/`).

References

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., and Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, 27(20):3968–3975.

Barthet, M. M., Moukarzel, K., Smith, K. N., Patel, J., and Hilu, K. W. (2015). Alternative translation initiation codons for the plastid maturase MatK: unraveling the pseudogene misconception in the Orchidaceae. *BMC Evolutionary Biology*, 15:210.

Bensasson, D., Zhang, D.-X., Hartl, D. L., and Hewitt, G. M. (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology and Evolution*, 16(6):314–321.

Berger, S. A., Krompass, D., and Stamatakis, A. (2011). Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology*, 60(3):291–302.

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., and De Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*, 29(6):358–367.

Boyce, K., Sievers, F., and Higgins, D. G. (2014). Simple chained guide trees give high-quality protein multiple sequence alignments. *Proceedings of the National Academy of Sciences USA*, 111(29):10556–10561.

Buhay, J. E. (2009). "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. *Journal of Crustacean Biology*, 29(1):96–110.

Calvignac, S., Konecny, L., Malard, F., and Douady, C. J. (2011). Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (numts). *Mitochondrion*, 11(2):246–254.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4):540–552.

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences USA*, 106(31):12794–12797.

Coissac, E., Riaz, T., and Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, 21(8):1834–1847.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319.

Dunning, L. T. and Savolainen, V. (2010). Broad-scale amplification of matK for DNA barcoding plants, a technical note. *Botanical Journal of the Linnean Society*, 164(1):1–9.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27(2):221–224.

Guindon, S., Delsuc, F., Dufayard, J.-F., and Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. *Methods in Molecular Biology*, 537:113–137.

Hebert, P. D., Cywinska, A., Ball, S. L., and Dewaard, J. R. (2003a). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1512):313–321.

Hebert, P. D., Ratnasingham, S., and De Waard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(suppl_1):S96–S99.

Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology*, 2(10).

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., et al. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16(10):1245–1257.

Kress, W. J. and Erickson, D. L. (2007). A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. *PLoS One*, 2(6).

Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PloS One*, 12(5):e0177459.

Lahaye, R., Van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T. G., and Savolainen, V. (2008). DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences USA*, 105(8):2923–2928.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278.

Leray, M. and Knowlton, N. (2015). Dna barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences USA*, 112(7):2076–2081.

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., and Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, 10(1):34.

Linard, B., Swenson, K., and Pardi, F. (2019). Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinformatics*, 35(18):3303–3312.

Liu, S., Li, Y., Lu, J., Su, X., Tang, M., Zhang, R., Zhou, L., Zhou, C., Yang, Q., Ji, Y., et al. (2013). SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution*, 4(12):1142–1150.

Lopez, J. V., Yuhki, N., Masuda, R., Modi, W., and O'Brien, S. J. (1994). Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution*, 39(2):174–190.

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1):538.

Moulton, M. J., Song, H., and Whiting, M. F. (2010). Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda: Insecta). *Molecular Ecology Resources*, 10(4):615–627.

Pompanon, F., Deagle, B. E., Symondson, W. O., Brown, D. S., Jarman, S. N., and Taberlet, P. (2012). Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, 21(8):1931–1950.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21):7188–7196.

Ramirez-Gonzalez, R., Yu, D. W., Bruce, C., Heavens, D., Caccamo, M., and Emerson, B. C. (2013). PyroClean: denoising pyrosequences from protein-coding amplicons for the recovery of interspecific and intraspecific genetic variation. *PLoS One*, 8(3).

Ranwez, V. and Chantret, N. (2020). Strengths and limits of multiple sequence alignment and filtering methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.2, pages 2.2:1–2.2:36. No commercial publisher | Authors open access book.

Ranwez, V., Chantret, N., and Delsuc, F. (2020). Aligning protein-coding nucleotide sequences with MACSE. To appear in Methods in Molecular Biology.

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., and Delsuc, F. (2018). MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, 35(10):2582–2584.

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PloS One*, 6(9).

Ratnasingham, S. and Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes*, 7(3):355–364.

Sarich, V. M. and Wilson, A. C. (1967). Immunological time scale for hominid evolution. *Science*, 158(3805):1200–1203.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541.

Scornavacca, C., Belkhir, K., Lopez, J., Dernat, R., Delsuc, F., Douzery, E. J. P., and Ranwez, V. (2019). OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Molecular Biology and Evolution*, 36(4):861–862.

Singh, T. R., Tsagkogeorga, G., Delsuc, F., Blanquart, S., Shenkar, N., Loya, Y., Douzery, E. J. P., and Huchon, D. (2009). Tunicate mitogenomics and phylogenetics: peculiarities of the Herdmania momus mitochondrial genome and support for the new chordate phylogeny. *BMC Genomics*, 10:534.

Smith, M. A., Fisher, B. L., and Hebert, P. D. (2005). DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1825–1834.

Sochat, V. V., Prybol, C. J., and Kurtzer, G. M. (2017). Enhancing reproducibility in scientific computing: Metrics and registry for Singularity containers. *PloS One*, 12(11):e0188511.

Song, H., Buhay, J. E., Whiting, M. F., and Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences USA*, 105(36):13486–13491.

Steinegger, M. and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028.

Stoeckle, M. Y. and Kerr, K. C. (2012). Frequency matrix approach demonstrates high sequence quality in avian BARCODEs and highlights cryptic pseudogenes. *PLoS One*, 7(8).

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8):2045–2050.

Taylor, C. A. and Knouft, J. H. (2006). Historical influences on genital morphology among sympatric species: gonopod evolution and reproductive isolation in the crayfish genus Orconectes (Cambaridae). *Biological Journal of the Linnean Society*, 89(1):1–12.

Taylor, H. and Harris, W. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, 12(3):377–388.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., and Hebert, P. D. (2005). Dna barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1847–1857.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences USA*, 74(11):5088–5090.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences USA*, 87(12):4576–4579.

Yang, C., Wang, X., Miller, J. A., de Blécourt, M., Ji, Y., Yang, C., Harrison, R. D., and Yu, D. W. (2014). Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators*, 46:379–389.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., and Ding, Z. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3(4):613–623.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8(2):357–366.

Chapter 2.4 Orthology: Definitions, Prediction, and Impact on Species Phylogeny Inference

Rosa Fernández[1]

Institute of Evolutionary Biology (Spanish National Research Council–University Pompeu Fabra), Barcelona, Spain
rosa.fernandez@ibe.upf-csic.es
https://orcid.org/0000-0002-4719-6640

Toni Gabaldón[2]

Barcelona Supercomputing Centre (BSC-CNS). Jordi Girona, 29. 08034. Barcelona, Spain
Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain
Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
toni.gabaldon.bcn@gmail.com
https://orcid.org/0000-0003-0019-1735

Christophe Dessimoz[3]

Department of Computational Biology, University of Lausanne, Switzerland
Center for Integrative Genomics, University of Lausanne, Switzerland
Centre for Life's Origins and Evolution, University College London, UK
Department of Computer Science, University College London, UK
Swiss Institute of Bioinformatics, Lausanne, Switzerland
Christophe.Dessimoz@unil.ch
https://orcid.org/0000-0002-2170-853X

Abstract

Orthology is a central concept in evolutionary and comparative genomics, used to relate corresponding genes in different species. In particular, orthologs are needed to infer species trees. In this chapter, we introduce the fundamental concepts of orthology relationships and orthologous groups, including some non-trivial (and thus commonly misunderstood) implications. Next, we review some of the main methods and resources used to identify orthologs. The final part of the chapter discusses the impact of orthology methods on species phylogeny inference, drawing lessons from several recent comparative studies.

How to cite: Rosa Fernández, Toni Gabaldón, and Christophe Dessimoz (2020). Orthology: definitions, inference, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 2.4, pp. 2.4:1–2.4:14. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

[1] RF was supported by a Marie Skłodowska-Curie fellowship (grant agreement 747607).
[2] TG acknowledges support from the European Union's Horizon 2020 research and innovation program under the grant agreement ERC-2016-724173, and from the Spanish Instituto Nacional de Bioinformática (INB) grant PT17/0009/0023 - ISCIII-SGEFI/ERDF.
[3] CD acknowledges support by Swiss National Science Foundation grant 183723.

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 2.4; pp. 2.4:1–2.4:14

A book completely handled by researchers.
No publisher has been paid.

1 Introduction

All life on earth shares a common origin. An evidence for this is, for instance, the existence of "universal" genes shared by all living beings. Indeed, we can find genes that are so similar within or between species that we can infer to be evolutionarily related and share ancestry–i.e. *homologous*–beyond reasonable doubt. Identifying homologous genes is of great interest, because it is the first step toward identifying what is conserved, and what has changed during evolution. In addition, because experimental characterisation of genes remains labour intensive, assessing evolutionary relationships provides a way to interpolate or extrapolate gene attributes among different species, such as the structure and function of the proteins they encode (Eisen 1998; Chapter 4.2 [Robinson-Rechavi 2020])–a goal for which understanding orthology relationships is key, as discussed below. .

One key refinement is to try to distinguish more precisely *how* homologous genes are related, giving rise to different homology subtypes. Homologs arising through speciation are called *orthologs* (Fitch, 1970); those arising through duplication are called *paralogs* (Fitch, 1970); those arising through whole genome duplication (also referred to as *homopolyploidization* or *autopolyploidization* in plants) as *ohnologs* (Leveugle et al., 2003); those through hybridization followed by genome doubling (*allopolyploidization*) are referred to as *homoeologs* (Huskins, 1931; Glover et al., 2016); those through lateral gene transfer as *xenologs* (Gray and Fitch, 1983).

Here, we focus mostly in orthologs, which are of particular importance in phylogenomics as they provide the basis to infer species phylogenies. In the first part, we review more precisely how orthology is defined and inferred. We start with orthology between two species, and then consider orthology in multispecies contexts. In the second part, we discuss the impact of orthology on phylogenetic inference.

2 Definitions, implications, complications

The term "ortholog" was coined by Walter Fitch nearly 50 years ago (Fitch, 1970):

> "It is not sufficient, for example, when reconstructing a phylogeny from amino acid sequences that the proteins be homologous. [...] there should be two subclasses of homology. Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism (for example, α and β hemoglobin) the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example α hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact). Phylogenies require orthologous, not paralogous, genes."

The definition was visionary, eloquent, and seemingly simple to grasp. Yet there are several implications and complications that have lead to frequent misunderstandings and inconsistencies in the literature. We consider these in turn.

Fitch defines orthology and paralogy as relationships between two genes, depending on the type of initial evolutionary event that gave rise to the pair. This implies that subsequent events, e.g. duplications of one and/or the other gene have no bearing on the type of relationship. Such duplications can however mean that a gene can have more than one orthologous counterpart in another species. In other words, orthology can be not only a one-to-one relationship, but also a one-to-many, many-to-one or many-to-many relationship.

Furthermore, note that the definitions are indifferent to the position of the genes on the genome. Consider e.g. a mammal gene retained in the human lineage and duplicated in the rodent lineage. Consider furthermore that one mouse copy has remained in its ancestral locus and the other one has moved elsewhere in the genome. Both rodent paralogous copies are orthologous to the human gene. To specify a conserved locus, the concept of "positional ortholog" has been proposed (Dewey, 2011).

In Fitch's examples, the paralogs ("α- and β-hemoglobin") belong to the same organism while the orthologs ("α-hemoglobin in man and mouse") belong to different species. Is paralogy still meaningful when the two genes are found in two different species? The answer is a resolute "yes". For instance, α-hemoglobin in mouse and β-hemoglobin in human are paralogs because they resulted from a duplication in a common ancestor of the two species.

A more tricky question is the converse: is orthology still meaningful if the two genes belong to the same species? To answer this, we need to consider the possibility that two genes resulting from a speciation event end up inside the same organism. This is unusual, but could happen through lateral gene transfer or hybridisation. However, in such cases, a different terminology is customarily used–xenologs or homoeologs respectively. Calling such genes "orthologs" would be consistent with Fitch's definition, but would be at odds with common usage in the literature.

2.1 From pairwise to groupwise orthology

Moving on beyond two genes, let us consider how orthology and paralogy apply to more than two species at a time. This generalisation is not a straightforward one because orthology and paralogy relationships are not transitive. That is, if gene A is orthologous to B, and B is orthologous to C, one cannot conclude that A and B are orthologous to each other. For instance, mouse has two insulin genes *Ins1* and *Ins2*, which duplicated within the rodent lineage (Shiao et al., 2008). Human has one copy, *INS*. Therefore, *Ins1* is orthologous to *INS, INS* is orthologous to *Ins2*, but *Ins1* is not orthologous but paralogous to *Ins2*. The same is true for paralogous relationships.

The original Fitch definition is that of a pairwise relationship between two genes. However, the number of pairwise relationships grows quadratically with the number of genes and species considered. Moreover, as we have seen, there is no straightforward extrapolation of pairwise relationships among groups of genes or across species. This, together with the difficulty of expressing and interpreting pairwise relationships when referring to groups of genes in several species, prompted the concept of "orthologous groups".

Two main kinds of orthologous groups have been proposed (Figure 1). One kind–which we refer to as "strict" orthologous groups–denotes sets of genes for which every two members are orthologous. This is the case for sets of one-to-one orthologs, as one-to-one orthology is transitive. More generally, it is also possible for such orthologous groups to span over duplication events as long as the resulting groups do not include any pair of paralogs. For instance, in the insulin example from above, a group containing *INS* and *Ins1* would fulfill this definition (Figure 1).

The other main kind of orthologous groups, called "hierarchical orthologous group" to avoid ambiguity, aims to identify sets of genes that have descended from a common ancestral gene in a given ancestral species. In the insulin example, since human *INS,* mouse *Ins1,* and mouse *Ins2* all descend from a common ancestral gene in the last common ancestor of all mammals, they are in a common hierarchical orthologous group at that level. By contrast, since *Ins1* and *Ins2* duplicated prior to the last murine common ancestor, these two genes are in different groups at the murine level (Figure 1). Thus, we can see that hierarchical

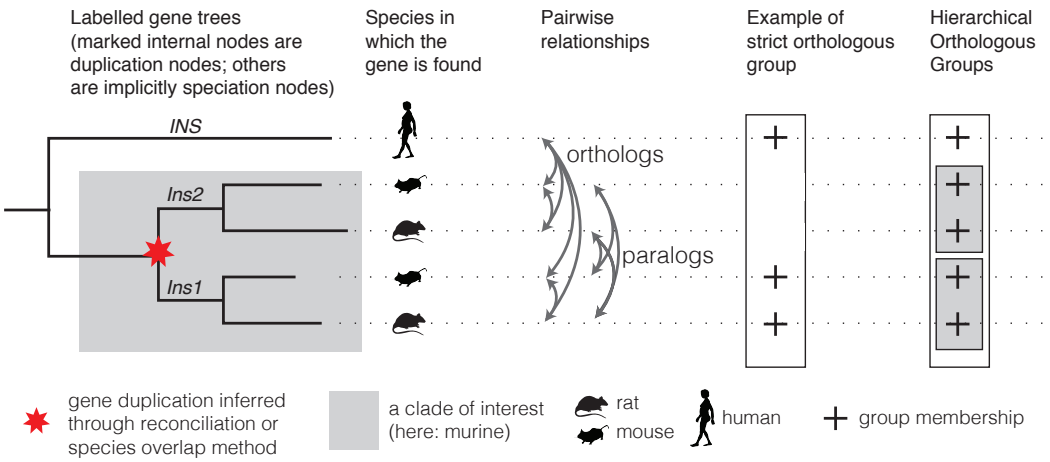


Figure 1 Conceptual overview of key concepts in this chapter. Gene tree with duplications and speciation nodes identified by reconciliation or species overlap; pairwise orthology and paralogy relationships; strict and hierarchical orthologous groups.

orthologous groups are defined with respect to specific clades. Furthermore, we can see their *hierarchical* nature in that groups defined with respect to deeper clades subsume multiple groups defined on their descendants–as the insulin example also illustrates. This definition of orthologous groups is similar to the older concept of "subfamily" used to describe a subset of members of a gene family that share a common ancestry (i.e. form a clade in the gene family tree).

One special type of hierarchical groups is worth mentioning. When dealing with two species only, the point of reference is implicitly meant to be their last common ancestor. In this case, hierarchical orthologous groups coincide with "ortholog clusters" as defined by the method *Inparanoid* (Remm et al., 2001).

For a more in-depth review of the different types of orthologous groups, we refer readers to Boeckmann et al. (2011).

2.2 Reconciled gene trees

As an alternative to orthologous groups, it is also possible to capture orthologous relationships using rooted gene trees which have their internal nodes labelled as speciation or duplication nodes (or possibly even more types of nodes). Such trees are commonly referred to as "labelled" or "reconciled" gene trees (Chapter 3.2 [Boussau and Scornavacca 2020]). All orthology and paralogy relationships among pairs of extant genes (i.e. pairs of leaves) in such trees can be deduced from the label associated with their last common ancestor: if the last common ancestor is a speciation node, the two genes are orthologs; if it is a duplication, they are paralogs (Figure 1). Methods to infer the duplication and speciation labels are reviewed below.

Likewise, hierarchical orthologous groups can be obtained from the clades rooted in the speciation nodes corresponding to the taxonomic range of interest. Therefore, labelled gene trees capture all orthology and orthologous group information. In addition, gene trees convey the order of gene duplications and quantify the amount of sequence divergence in the branch lengths. Now that we have defined orthology, paralogy, orthologous groups, and reconciled trees, we turn to methods to infer these various types of relationships.

3 State-of-the-art methods and resources

In this section, we provide an overview of methods and databases for orthology inference (Table 1). Methods are commonly divided into two main groups: tree-based and graph-based methods. Tree-based methods, as the name indicates, explicitly infer gene trees at some stage of their algorithms. By contrast, graph-based methods avoid inferring trees and instead compare sequences in a pairwise fashion, and build a graph with genes as vertices and some measure of sequence similarity as edges. We detail the two types of approach in turn, and refer to some of the popular associated algorithms and databases.

3.1 Tree-based approaches

As we already discussed in the definition section, tree-based orthology inference methods reconstruct a gene tree for a group of homologous sequences to then infer the type of evolutionary event represented by each internal node of the tree. To infer events at internal nodes, the conventional approach is to perform "gene tree/species tree reconciliation". This can be done in a parsimony or in a likelihood framework (see Boussau et al. 2013; Chapter 3.2 [Boussau and Scornavacca 2020]). Alternatively, the labelling of internal nodes can be determined by the method of species overlap, which labels as duplication node any internal node which has the same species represented in more than one of its child subtrees (van der Heijden et al., 2007; Huerta-Cepas et al., 2007). Thus the species overlap approach does not require or assume any species tree. Or, to be more precise, it considers a fully-unresolved species tree. Hence, it relies on the two copies that result from each gene duplication to be retained in at least one species, which is often the case in practice. This algorithm is considerably more robust to topological diversity in the gene trees–in contrast to conventional gene/species tree reconciliation methods, which tend to introduce duplication events to explain any departure from the canonical species tree.

Several resources provide reconciled gene trees. PhylomeDB (Huerta-Cepas et al., 2014) and MetaPhOrs (Pryszcz et al., 2011) use the species overlap approach. For each reference species in the database, PhylomeDB infers a gene tree starting from each protein (each "seed"), and refers to the resulting set of trees as the *phylome* of that species. The species overlap method is also used and is available as part of the ETE software library (Huerta-Cepas et al., 2016a). Species overlap is extended to unrooted gene trees with UPhO (Ballesteros and Hormiga, 2016). PANTHER trees (Mi et al., 2017) infers reconciliation for all PANTHER families using the GIGA algorithm (Thomas, 2010), a gene/species reconciliation method. Likewise EnsemblCompara infers reconciled gene trees relating all Ensembl genomes using the TreeBeST algorithm (Vilella et al., 2008).

3.2 Graph-based approaches

Graph-based approaches are based on comparisons between pairs of genes within and between species. They are all based on the observation that, for pairs of genes between two species, orthologs tend to be the pairs of sequences that have diverged the least (Wolf and Koonin, 2012; Dalquen and Dessimoz, 2013). This is because until the speciation event that relates the two species, the orthologs were the same genes, while paralogs are the result of earlier duplications, and thus have more time to diverge.

This insight gave rise to the first large-scale orthology prediction approach, the basic "bidirectional best hit" (BBH) approach (Overbeek et al., 1999), which considers the pairs with mutually highest alignment scores, or its phylogenetic distance-based counterpart entitled

Table 1 Orthology methods mentioned in this chapter. For more methods, consult the *Quest for Orthologs* consortium website at `https://questfororthologs.org/orthology_databases` as well as Chapter 3.2 [Boussau and Scornavacca 2020] on reconciliation methods.

Method	Type	Comments	Ref.
BUSCO	Graph	Based on precomputed "universal single-copy" genes (defined for a number of standard clades), and thus inherently limited to these. Originally developed to assess genome completeness.	(Waterhouse et al., 2017)
COG/KOG	Graph	One of the first methods, still widely used for prokaryotic data. Includes a manual curation step.	(Tatusov et al., 2003)
EggNOG	Hybrid	Originally developed as extension of COG/KOG. Recent versions also include tree-based refinements.	(Huerta-Cepas et al., 2016b)
ETE 3.0	Tree	General purpose tree analysis and visualisation package for Python, with species overlap function.	(Huerta-Cepas et al., 2016a)
GIGA	Tree	Gene/species tree reconciliation algorithm used in the PANTHER database. Also includes a heuristic for lateral gene transfer detection.	(Thomas, 2010)
HaMSTR	Graph	The method uses a reference species to define one Hidden Markov Model per orthologous group, followed by reciprocal best hit within a family	(Ebersberger et al., 2009)
Hieranoid	Graph	Successor of Inparanoid to infer hierarchical orthologous groups from multiple species	(Kaduk et al., 2017)
Inparanoid	Graph	Infers orthologous groups independently for each pair of species.	(Sonnhammer and Östlund, 2015)
MetaPhOrs	Hybrid	Meta-method integrating predictions from multiple sources.	(Pryszcz et al., 2011)
OMA	Graph	Infers both types of groups reviewed in this chapter: strict groups (suitable as markers for species tree inference) and hierarchical orthologous groups.	(Altenhoff et al., 2018)
OrthoDB	Graph	Infers hierarchical orthologous groups. Used to infer the single-copy universal gene models of BUSCO.	(Zdobnov et al., 2017)
OrthoFinder	Graph	Infers hierarchical orthologous group with respect to the deepest speciation level only (the last common ancestor)	(Emms and Kelly, 2015)
OrthoInspector	Graph	Provides phylogenetic profiles as well.	(Nevers et al., 2019)
OrthoMCL	Graph	Groups inferred by OrthoMCL do not have a straightforward interpretation (they are neither strict nor hierarchical). Often used in combination with other methods and/or criteria.	(Li et al., 2003)
PhylomeDB	Tree	Based on species overlap method.	(Huerta-Cepas et al., 2014)
UPhO	Tree	Species overlap method considering multiple gene tree rootings.	(Ballesteros and Hormiga, 2016)

"reciprocal shortest distance" (RSD) (Wall et al., 2003).

However, BBH and RSD do not deal well with many-to-many orthology relationships, resulting in missing pairs (Dalquen and Dessimoz, 2013). To address this, the Inparanoid algorithm provided a way to identify many-to-many orthology relationships (Remm et al., 2001). Furthermore, BBH and RSD can fail in case of differential gene loss–a situation where the corresponding ortholog is simply missing in both species, resulting in paralogs being wrongly identified as orthologs. The OMA algorithm introduced the use of third-party species, which might have retained both copies, which could thus act as "witnesses of non-orthology" (Dessimoz et al., 2006).

The other limitation of BBH and RSD is that they do not obviously generalise to groupwise orthology. The COGs database pioneered the use of "triangles" of pairwise orthologs (complemented by manual curation) to build multi-species orthologous groups (Tatusov et al., 1997). OrthoMCL used Markov clustering instead (Li et al., 2003). One issue with OrthoMCL is however that the granularity of the resulting groups depends on the choice of parameter ("inflation parameter"), which makes it harder to interpret the results.

The main graph-based resources include EggNOG (Huerta-Cepas et al., 2016b), HaMStR (Ebersberger et al., 2009), Inparanoid/HieranoiDB (Sonnhammer and Östlund, 2015; Kaduk et al., 2017), OMA (Altenhoff et al., 2018), OrthoDB (Zdobnov et al., 2017), OrthoFinder (Emms and Kelly, 2015), and OrthoInspector (Nevers et al., 2019).

4 Impact on phylogenomic inference: resolving the Tree of Life

Resolving the Tree of Life has been one of the prevailing questions in evolutionary biology at all systematic levels since the origin of phylogenetics. From bacteria to eukaryotes, from archaea to metazoa, great scientific efforts have been devoted towards understanding the evolutionary relationships between organisms.

The first sources of phylogenetic information to infer species trees were morphological characters. These characters were first classified as homologous or not based on taxonomic comparisons, then into ancestral or derived; finally phylogenetic interrelationships were inferred based on a parsimony criterium (Fitch, 1971). With the advent of molecular biology techniques, scientific efforts shifted largely to the use of molecular markers, which were aligned, concatenated (if several markers were used) and used to reconstruct a phylogeny. This approach would only provide sensible results if aligned sequences are orthologous to each other, as orthologs define speciation nodes, which constitute the only type of nodes that are expected in a species tree. If some of the sequences included have paralogous relationships, then some of the reconstructed nodes will indeed represent duplications and the resulting topology will be faulty with respect to the aimed species tree.

Initially, the experimental design in molecular phylogenetics included the identification of highly conserved regions in the organismal lineage of interest, that were amplified with specific probes by means of a polymerase chain reaction (PCR). As the same marker gene –i.e. the orthologous gene–was specifically sequenced from each of the species of interest, there was no need to search for orthologs. However, problems such as cross-amplification of paralogs, non-specific amplifications in the absence of the ortholog, or hidden paralogy issues, were common problems that could complicate the process of species tree reconstruction and have their root in the failure of obtaining a fully orthologous sequence dataset. With the advent of high-throughput sequencing and the availability of complete (or nearly complete) genomes and transcriptomes, one can in principle choose among virtually any marker gene. In these cases, there is a need of inferring orthologous genes from the source genomic datasets,

and doing so correctly is pivotal for accurately reconstructing a species tree (see Chapter 2.1 [Simion et al. 2020]). As we will see below, despite the availability of automated methods, problems are likely to be encountered.

The past decade has seen an explosion of genome and transcriptome sequences from non-model organisms. More often than not, phylogenomic datasets include transcriptomes and low-coverage genomes that are incomplete, and contain errors and unresolved isoforms. These characteristics can severely violate the assumptions underlying some orthology inference methods. As a result, different orthology methods can result in very different phylogenetic inferences. Despite this fact, the effect of orthology inference is not commonly considered in typical phylogenomic analyses aimed at reconstructing species trees. Instead, methodological discussions have largely focused on the effect of phylogenetic reconstruction parameters such as the chosen models of substitution applied to the datasets, or on the effect of confounding factors, including missing data, compositional heterogeneity, or incomplete taxon sampling, among others. This is perhaps best epitomized by the intense debate around the position of ctenophores (Dunn et al., 2008; Hejnol et al., 2009; Moroz et al., 2014; Borowiec et al., 2015; Whelan et al., 2015; Shen et al., 2017) or sponges (Philippe et al., 2009; Pick et al., 2010; Philippe et al., 2011; Pisani et al., 2015; Simion et al., 2017) as the earliest-branching phylum in the Animal Tree of Life (see Chapter 2.1 [Simion et al. 2020]).

Orthology benchmarking requires curated information about the underlying gene and species trees (e.g., the *Quest for Orthologs* benchmark service [Altenhoff et al. 2016]). As a consequence, when the goal is to infer the species tree, a comparison of orthology inference methods (everything else in the analytical pipeline being unchanged) appears as the most appropriate alternative to assess the robustness of the resulting topology. Yet few studies have compared how sets of orthologs inferred through different methods vary and how it affects species tree reconstruction. Shen *et al.* (Shen et al., 2018) compared the performance of OrthoMCL (Li et al., 2003) refined with PhyloTreePruner (Kocot et al., 2013), BUSCO v.2.0.1 (Waterhouse et al., 2017) and PhylomeDB v4 (Huerta-Cepas et al., 2014) in a data set composed of 332 budding yeast (Saccharomycotina) genomes. They compared the overlap between the refined OrthoMCL orthologous groups (with a size of 2,408 orthologous groups, referred to as OGs hereafter) with the BUSCO and PhylomeDB ones, respectively. From a total of 1,292 BUSCO OGs, a large majority were recovered by OrthoMCL as well (1,081 OGs). However, OrthoMCL recovered less than half of the PhylomeDB OGs (819 out of 1,838). Overall, the resulting topologies after the analysis of the concatenated data sets differed in 10% of the nodes (32 out of 331 nodes).

In two studies dealing with spiders interrelationships (a much shallower systematic level than the previous example), Fernández *et al.* (Fernández et al., 2018) and Kallal *et al.* (Kallal et al., 2018) compared OGs inferred by BUSCO v1.1b (Simão et al., 2015) and UPhO (Ballesteros and Hormiga, 2016). Contrarily to the example of Shen et al. (Shen et al., 2018), these authors did not analyse the matrix resulting from the intersection of both orthology inference methods, but the BUSCO OGs and UPhO OGs individually. Both studies found congruence between most of the analyses in the concatenated matrices, with minimal topological effects from orthology assessment despite recovering an overlap of as low as 4.3% of OGs between both methods, as reported in Kallal et al. (2018).

Finally, Altenhoff et al. (2019) compared several orthology methods (OMA, OrthoMCL, OrthoFinder, HaMStR, and BUSCO) on a reconstruction of the Lophotrochozoa phylogeny. The number of orthologous groups recovered varied quite substantially–ranging from 384 (BUSCO) to 2162 (OMA). Furthermore, the accuracy and branch support of Bayesian and Maximum Likelihood trees reconstructed from these groups varied considerably, suggesting

that for difficult phylogenies such as Lophotrochozoa, the choice of orthology inference method can lead to different conclusions.

All in all, while Fernández et al. (2018) and Kallal et al. (2018) found congruence between most topologies despite differences in the sequences used to infer the species trees–therefore suggesting strong signal in the data robust to differences in orthology inference– Shen et al. (2018) and Altenhoff et al. (2019) found that as much as 10% of the nodes varied between topologies. These results highlight the importance of comparing orthology inference methods in each data set as they may strongly affect the resulting species tree topology.

The selection of a proper orthology inference software is of particular importance in complex evolutionary scenarios where gene and genome duplications are frequent, as is the case in plants. Orthology inference methods developed without explicit consideration for such duplication events, such as OrthoMCL (Li et al., 2003), have been reported to be potentially problematic in plants because they tend to break gene families apart instead of retaining its structure (McKain et al., 2018). Instead, other methods better able to account for gene duplications have been recommended in this challenging phylogenomic scenarios, for example OrthoFinder (Emms and Kelly, 2015), OMA (Altenhoff et al., 2018), PhylomeDB (Huerta-Cepas et al., 2014) or all-by-all BLAST followed by Markov clustering and tree-based orthology pruning (McKain et al., 2018; Yang and Smith, 2014)

Regardless of the software selected for orthology inference, the inclusion of paralogous sequences may result in different outcomes. In some cases, such as in shallow-level phylogenies (e.g., at the level of order, genus, etc.), species tree reconstruction may not be affected by paralogs as far as they are recent enough to be monophyletic for each lineage. In other cases, paralogs have been even proven useful as additional loci for phylogenetics, as reads from the two paralogous sequences can be sorted and assembled into separate, orthologous alignments when the relative age of a genome duplication is known (Johnson et al., 2016).

5 Conclusions

As we have reviewed in this chapter, orthology is a fundamental concept for phylogenomics. The terminology, its implications, and the daunting array of methods led to some confusion in the early days of genomics. This has noticeably improved, in large part thanks to a sustained community effort around the *Quest for Orthologs* consortium (Gabaldón et al., 2009; Dessimoz et al., 2012; Sonnhammer et al., 2014; Forslund et al., 2017; Glover et al., 2019).

Yet challenges remain. In the context of more than two species, the concept of an orthologous group remains often imprecise in the literature; we have yet to attain the same level of understanding for groupwise orthology as for pairwise orthology. Comparisons among methods has also mainly focused on pairwise orthology. But phylogenomic tree inference requires groups, and several recent studies have observed substantial differences in the trees obtained from different orthologous group reconstruction techniques. Thus, to resolve difficult phylogenies, it may be necessary to better understand and characterise the impact of orthology on tree inference.

References

Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., Forslund, K., Huerta-Cepas, J., Linard, B., Pereira, C., Pryszcz, L. P., Schreiber, F., da Silva, A. S., Szklarczyk, D., Train, C.-M., Bork, P., Lecompte, O., von Mering, C.,

Xenarios, I., Sjölander, K., Jensen, L. J., Martin, M. J., Muffato, M., Quest for Orthologs consortium, Gabaldón, T., Lewis, S. E., Thomas, P. D., Sonnhammer, E., and Dessimoz, C. (2016). Standardized benchmarking in the quest for orthologs. *Nat. Methods*, 13(5):425–430.

Altenhoff, A. M., Glover, N. M., Train, C.-M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., de Farias, T. M., Zile, K., Stevenson, C., Long, J., Redestig, H., Gonnet, G. H., and Dessimoz, C. (2018). The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, 46(D1):D477–D485.

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Warwick Vesztrocy, A., Dalquen, D. A., Müller, S., Telford, M. J., Glover, N. M., Dylus, D., and Dessimoz, C. (2019). OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.*, 29(7):1152–1163.

Ballesteros, J. A. and Hormiga, G. (2016). A new orthology assessment method for phylogenomic data: Unrooted phylogenetic orthology. *Mol. Biol. Evol.*, 33(9):2481.

Boeckmann, B., Robinson-Rechavi, M., Xenarios, I., and Dessimoz, C. (2011). Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief. Bioinform.*, 12(5):423–435.

Borowiec, M. L., Lee, E. K., Chiu, J. C., and Plachetzki, D. C. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the ctenophora as sister to remaining metazoa. *BMC Genomics*, 16:987.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Boussau, B., Szöllosi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Res.*, 23(2):323–330.

Dalquen, D. a. and Dessimoz, C. (2013). Bidirectional best hits miss many orthologs in Duplication-Rich clades such as plants and animals. *Genome Biol. Evol.*, 5(10):1800–1806.

Dessimoz, C., Boeckmann, B., Roth, A. C. J., and Gonnet, G. H. (2006). Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res.*, 34(11):3309–3316.

Dessimoz, C., Gabaldón, T., Roos, D. S., Sonnhammer, E. L. L., Herrero, J., and Quest for Orthologs Consortium (2012). Toward community standards in the quest for orthologs. *Bioinformatics*, 28(6):900–904.

Dewey, C. N. (2011). Positional orthology: putting genomic evolutionary relationships into context. *Brief. Bioinform.*, 12(5):401–412.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sørensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., and Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–749.

Ebersberger, I., Strauss, S., and von Haeseler, A. (2009). HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol.*, 9:157.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, 8(3):163–167.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, 16:157.

Fernández, R., Kallal, R. J., Dimitrov, D., Ballesteros, J. A., Arnedo, M. A., Giribet, G., and Hormiga, G. (2018). Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. *Curr. Biol.*, 28(13):2190–2193.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.*, 19(2):99–113.

Fitch, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.*, 20(4):406–416.

Forslund, K., Pereira, C., Capella-Gutierrez, S., Sousa da Silva, A., Altenhoff, A., Huerta-Cepas, J., Muffato, M., Patricio, M., Vandepoele, K., Ebersberger, I., Blake, J., Fernández Breis, J. T., Quest for Orthologs Consortium, Boeckmann, B., Gabaldón, T., Sonnhammer, E., Dessimoz, C., and Lewis, S. (2017). Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics*.

Gabaldón, T., Dessimoz, C., Huxley-Jones, J., Vilella, A. J., Sonnhammer, E. L., and Lewis, S. (2009). Joining forces in the quest for orthologs. *Genome Biol.*, 10(9):403.

Glover, N., Dessimoz, C., Ebersberger, I., Forslund, S. K., Gabaldón, T., Huerta-Cepas, J., Martin, M.-J., Muffato, M., Patricio, M., Pereira, C., da Silva, A. S., Wang, Y., Sonnhammer, E., and Thomas, P. D. (2019). Advances and applications in the quest for orthologs. *Mol. Biol. Evol.*, 36(10):2157–2164.

Glover, N. M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: What are they and how do we infer them? *Trends Plant Sci.*, 21(7):609–621.

Gray, G. S. and Fitch, W. M. (1983). Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from staphylococcus aureus. *Mol. Biol. Evol.*, 1(1):57–66.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., Martinez, P., Baguñà, J., Bailly, X., Jondelius, U., Wiens, M., Müller, W. E. G., Seaver, E., Wheeler, W. C., Martindale, M. Q., Giribet, G., and Dunn, C. W. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. Biol. Sci.*, 276(1677):4261–4270.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M., and Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, 42(Database issue):D897–902.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldón, T. (2007). The human phylome. *Genome Biol.*, 8(6):R109.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016a). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, 33(6):1635–1638.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C., and Bork, P. (2016b). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, 44(D1):D286–93.

Huskins, C. L. (1931). A cytological study of vilmorin's unfixable dwarf wheat. *J. Genet.*, 25(1):113–124.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega, N. J. C., and Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.*, 4(7).

Kaduk, M., Riegler, C., Lemp, O., and Sonnhammer, E. L. L. (2017). HieranoiDB: a database of orthologs inferred by hieranoid. *Nucleic Acids Res.*, 45(D1):D687–D690.

Kallal, R. J., Fernández, R., Giribet, G., and Hormiga, G. (2018). A phylotranscriptomic backbone of the orb-weaving spider family araneidae (arachnida, araneae) supported by multiple methodological approaches. *Mol. Phylogenet. Evol.*, 126:129–140.

Kocot, K. M., Citarella, M. R., Moroz, L. L., and Halanych, K. M. (2013). PhyloTreePruner: A phylogenetic Tree-Based approach for selection of orthologous sequences for phylogenomics. *Evol. Bioinform. Online*, 9:429–435.

Leveugle, M., Prat, K., Perrier, N., Birnbaum, D., and Coulier, F. (2003). ParaDB: a tool for paralogy mapping in vertebrate genomes. *Nucleic Acids Res.*, 31(1):63–67.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13(9):2178–2189.

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., and Yang, Y. (2018). Practical considerations for plant phylogenomics. *Appl. Plant Sci.*, 6(3):e1038.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2017). PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, 45(D1):D183–D189.

Moroz, L. L., Kocot, K. M., Citarella, M. R., Dosung, S., Norekian, T. P., Povolotskaya, I. S., Grigorenko, A. P., Dailey, C., Berezikov, E., Buckley, K. M., Ptitsyn, A., Reshetov, D., Mukherjee, K., Moroz, T. P., Bobkova, Y., Yu, F., Kapitonov, V. V., Jurka, J., Bobkov, Y. V., Swore, J. J., Girardo, D. O., Fodor, A., Gusev, F., Sanford, R., Bruders, R., Kittler, E., Mills, C. E., Rast, J. P., Derelle, R., Solovyev, V. V., Kondrashov, F. A., Swalla, B. J., Sweedler, J. V., Rogaev, E. I., Halanych, K. M., and Kohn, A. B. (2014). The ctenophore genome and the evolutionary origins of neural systems. *Nature*, 510(7503):109–114.

Nevers, Y., Kress, A., Defosset, A., Ripp, R., Linard, B., Thompson, J. D., Poch, O., and Lecompte, O. (2019). OrthoInspector 3.0: open portal for comparative genomics. *Nucleic Acids Res.*, 47(D1):D411–D418.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.*, 96(6):2896–2901.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.*, 9(3):e1000602.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D. J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G., and Manuel, M. (2009). Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.*, 19(8):706–712.

Pick, K. S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D. J., Wrede, P., Wiens, M., Alié, A., Morgenstern, B., Manuel, M., and Wörheide, G. (2010). Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol. Biol. Evol.*, 27(9):1983–1987.

Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., and Wörheide, G. (2015). Genomic data do not support comb jellies as the sister group to all other animals. *Proc. Natl. Acad. Sci. U. S. A.*, 112(50):15402–15407.

Pryszcz, L. P., Huerta-Cepas, J., and Gabaldón, T. (2011). MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.*, 39(5):e32.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, 314(5):1041–1052.

Robinson-Rechavi, M. (2020). Molecular evolution and gene function. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.2, pages 4.2:1–4.2:20. No commercial publisher | Authors open access book.

Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol*, 1(5):126.

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., Haase, M. A. B., Wisecaver, J. H., Wang, M., Doering, D. T., Boudouris, J. T., Schneider, R. M., Langdon, Q. K., Ohkuma, M., Endoh, R., Takashima, M., Manabe, R.-I., Čadež, N., Libkind, D., Rosa, C. A., DeVirgilio, J., Hulfachor, A. B., Groenewald, M., Kurtzman, C. P., Hittinger, C. T., and Rokas, A. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. *Cell*, 0(0).

Shiao, M.-S., Liao, B.-Y., Long, M., and Yu, H.-T. (2008). Adaptive evolution of the insulin two-gene system in mouse. *Genetics*, 178(3):1683–1691.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., and Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.*, 27(7):958–967.

Sonnhammer, E. L. L., Gabaldón, T., Sousa da Silva, A. W., Martin, M., Robinson-Rechavi, M., Boeckmann, B., Thomas, P. D., Dessimoz, C., and Quest for Orthologs consortium (2014). Big data and other challenges in the quest for orthologs. *Bioinformatics*, 30(21):2993–2998.

Sonnhammer, E. L. L. and Östlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, 43(Database issue):D234–9.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Sridhar Rao, B., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. a. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, 278(5338):631–637.

Thomas, P. D. (2010). GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics*, 11:312.

van der Heijden, R. T. J. M., Snel, B., van Noort, V., and Huynen, M. a. (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*, 8:83.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2008). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19(2):327–335.

Wall, D. P., Fraser, H. B., and Hirsh, a. E. (2003). Detecting putative orthologs. *Bioinformatics*, 19(13):1710–1711.

2.4:14 REFERENCES

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., and Zdobnov, E. M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*

Whelan, N. V., Kocot, K. M., Moroz, L. L., and Halanych, K. M. (2015). Error, signal, and the placement of ctenophora sister to all other animals. *Proc. Natl. Acad. Sci. U. S. A.*, 112(18):5773–5778.

Wolf, Y. I. and Koonin, E. V. (2012). A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol. Evol.*, 4(12):1286–1294.

Yang, Y. and Smith, S. A. (2014). Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.*, 31(11):3081–3092.

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simão, F. A., Ioannidis, P., Seppey, M., Loetscher, A., and Kriventseva, E. V. (2017). OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.*, 45(D1):D744–D749.

Chapter 2.5 Ancestral Genome Organization as a Diagnosis Tool for Phylogenomics

Eric Tannier
INRIA and Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558,
[Villeurbanne, France]
eric.tannier@inria.fr
https://orcid.org/0000-0002-3681-7536

Adelme Bazin
Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay
[91057 Evry, France]
adelme.bazin@genoscope.cns.fr

Adrián A. Davín
RIKEN Center for Advanced Intelligence Project (AIP)
[36-1 Yoshida Honmachi, Sakyo-ku, Kyoto, Japan]
adrian.arellanodavin@riken.jp
https://orcid.org/0000-0003-4945-4938

Laurent Guéguen
Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558
[Villeurbanne, France]
laurent.gueguen@univ-lyon1.fr

Sèverine Bérard
ISEM, Université de Montpellier, CNRS, IRD, EPHE,
[Montpellier, France]
severine.berard@univ-montellier.fr

Cédric Chauve[1]
Department of Mathematics, Simon Fraser University,
[8888 University Drive, Burnaby (BC), V5A 1S6, Canada]
LaBRI, Université de Bordeaux
[351 Cours de la Libération, 33405 Talence Cedex, France]
cedric.chauve@sfu.ca
https://orcid.org/0000-0001-9837-1878

Abstract

The reconstruction of the chromosomal organization of ancient genomes has many applications in comparative and evolutionary genomics. Here we propose a novel, methodological, use for these predicted ancestral syntenies, directly focused on phylogenomics. It is a way to assess the accuracy of gene trees and species trees. We use a method that reconstructs, from gene trees and extant gene orders, ancestral adjacencies, i.e. the immediate neighborhood between pairs of genes, independently for each pair. This independence allows to split the computations into many independent problems that can each be solved exactly using efficient algorithms, but might result in sets of ancestral adjacencies that are incompatible with the expected linear or circular structure of chromosomes. We show here that this drawback can actually be turned into a useful

[1] This work benefited from the support of ComputeCanada.

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 2.5; pp. 2.5:1–2.5:19
A book completely handled by researchers.
No publisher has been paid.

feature. We show on simulated data that the degree of linearity of the reconstructed ancestral gene orders is well correlated to the accuracy of the input gene trees. Moreover, a localized error in the species trees results in a burst of non linearity of ancestral genomes at the wrong node. We eventually show that integrated phylogenomic methods expectedly lead to better linearity scores than methods based on gene alignments only. Allowing a method to output an unrealistic result, but proving that the expected output is closer to realistic when the input is closer to correct, we thus provide an original validation protocol for standard evolutionary studies.

How to cite: Eric Tannier, Adelme Bazin, Adrián A. Davín, Laurent Guéguen, Sèverine Bérard, and Cédric Chauve (2020). Ancestral Genome Organization as a Diagnosis Tool for Phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 2.5, pp. 2.5:2–2.5:19. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Supplement Material `https://github.com/sberard/SAGe`

1 Introduction

Rearrangements of gene organization along chromosomes were discovered long before the molecular structure of DNA (Sturtevant, 1921). The comparison of genetic maps or polytene chromosomes were seen in the first half of the XXth century as a promising approach to reconstruct evolutionary relationships or ancestral configurations (Babcock and Navashin, 1930; Dobzhansky and Sturtevant, 1938), and advance the knowledge on extant and extinct biodiversity. This later motivated the development of genetics, cytogenetics, or bioinformatics techniques aimed at detecting "structural" mutations of chromosomes (Timoshevskiy et al., 2013). The definition of what is a structural mutation largely depends on the observation technique used to detect them. It can consider every mutation involving several contiguous nucleotides (starting with small indels, micro-satellites or micro-inversions), or be limited to very large-scale mutations that affect gene content and orders, such as large inversions or chromosomal translocations (that can be detected by genetics or cytogenetics techniques). As genes are often taken as evolutionary units by bioinformatics and phylogenomics methods, genome rearrangements are usually limited to structural mutations whose breakpoints are located in non-coding regions and that change the organisation of genes along the genome.

Rearrangements have an important role in several evolutionary processes such as adaptation, speciation, sex differentiation, polyploidization (Fuller et al., 2018; Lemaitre et al., 2009). Knowing ancestral configurations can thus inform on conserved structures, functional gene clusters (Abrouk et al., 2010), as well as on patterns and processes of the history of wild or domestic biodiversity (Murat et al., 2012).

1.1 Ancestral gene order reconstruction methods

The reconstruction of ancient genome organization has been called *paleogenomics*, a term shared with ancient genome sequencing (Pont et al., 2019).

Over the last 25 years, there has been an intense research activity in developing computational methods for the reconstruction of ancestral gene orders, that we reviewed extensively by Anselmetti et al. (2018b). We distinguish two main families of methods: chromosome based methods and adjacency based methods.

Chromosome based methods take as input a species phylogeny, the gene orders of extant species in this phylogeny, and a genome rearrangement evolutionary model. Their aim is to infer ancestral gene orders and evolutionary scenarios along the branches of the species phylogeny. Ancestral orders and scenarios can be given a score (parsimony, likelihood) according to the model, and methods can optimize or sample according to this score.

Such methods are natural extensions to gene orders of ancestral sequence reconstruction methods, reviewed for example by Groussin et al. (2016); Joy et al. (2016). However, unlike ancestral sequence reconstruction, ancestral gene order reconstruction is computationally intractable for almost all genome rearrangement models. Indeed, even if gene scale evolution events as duplication and loss are ignored, and if there are only 3 species in the species tree, the parsimony problem is NP-complete (see Tannier et al., 2009; Kovác, 2014, and references there). If duplications are allowed, even the problem of computing the pairwise distance between two gene orders is NP-complete (e.g. see Blin et al., 2007; Angibaud et al., 2009).

To skirt this computational challenge, adjacency based methods model the evolution of the physical link between two consecutive genes only, called *adjacencies*, instead of full chromosomes. In this framework, ancestral adjacencies are reconstructed along the species phylogeny from the pattern of presence/absence of extant adjacencies, using a model allowing gains and losses of adjacencies. The set of inferred ancestral adjacencies for a specific ancestral species then forms an *adjacency graph* whose vertices are the ancestral genes and edges the ancestral gene adjacencies. A side effect of inferring ancestral adjacencies using such an approach that considers the evolution of each adjacency independently from the others is that the adjacency graph may not have the expected structure of a chromosome, that is, a collection of paths and cycles. In order to present a structure compatible with a set of chromosomes, or at least scaffolds, some methods select a subset of the inferred ancestral adjacencies which form a collection of paths and/or cycles. Computationally more tractable, adjacency based approaches can handle unequal gene content and gene duplication, gain and loss, and several methods have been developed that allow ancestral gene orders to contain duplicated genes (Ma et al., 2008; Rajaraman and Ma, 2016; Zhou et al., 2017).

Here, we consider again such methods, but from another point of view: we make the hypothesis that syntenic conflicts might be caused by errors in the earlier steps of the whole pipeline, especially the construction of the reconciled gene trees (see Chapter 3.2 [Boussau and Scornavacca 2020]). This hypothesis has been considered in several studies (Boussau et al., 2013; Peres and Crollius, 2015; Duchemin et al., 2017; Anselmetti et al., 2018a; Zerbino et al., 2018), but never assessed through experiments. In this paper, we provide a first proof of principle that indeed syntenic conflict in reconstructed ancestral gene orders can be correlated to errors in earlier steps of a phylogenomics pipeline.

1.2 Impact of errors on the linearity of reconstructed genomes

As described above, ancestral gene orders are typically obtained at the end of a multi-step sequential phylogenomics pipeline starting with genome assemblies and leading to the inference of ancestral gene adjacencies, which link consecutive genes in ancestral chromosomes, and ultimately ancestral gene orders. Intermediate steps include gene annotation (Chapter 4.1 [Necsulea 2020]), gene clustering into gene families (Chapter 2.4 [Fernández et al. 2020]), multiple sequence alignment of genes within families (Chapter 2.2 [Ranwez and Chantret 2020]), gene tree and species tree reconstruction (Chapters 1.2 and 1.4 [Stamatakis and Kozlov 2020; Lartillot 2020]) and gene tree reconciliation (Chapter 3.2 [Boussau and Scornavacca 2020]). Each step in such pipelines is susceptible to introduce errors that can propagate further in the pipeline. For example, the effect of errors in multiple alignments for species

tree reconstructions has been explored by Philippe et al. (2017), and the effect or errors in gene tree for reconciliations has been investigated by Hahn (2007). The effect of model choice has recently been investigated by Hoff et al. (2016); Yang and Zhu (2018), as well as the effect of the phylogenetic software choice (Zhou et al., 2018). And even bugs in many standard software can blur the results (Czech et al., 2017). Along these lines of research, we propose to investigate the effect of errors in gene trees and species trees on the *linearity* (or more precisely non-linearity) observed in ancestral gene adjacencies, and propose to use the latter to correct phylogenetic trees.

The notion of linearity is related to the arrangement of genes along chromosomes as defined by ancestral gene adjacencies. In this work we assume that a genome is composed of a set of linear and/or circular molecules carrying genes – chromosomes, organelles, plasmids, . . . – with genes *totally ordered* along each molecule. This implies that a correct adjacency graph, representing an actual gene order, whether it is extant or ancestral, is a collection of paths and/or cycles. This is an approximation as it is common in extant genomes that genes overlap, or are included one in an other, or that the definition of their limits lack precision due to alternative splicing for example. Nevertheless, a vast majority of genes in cellular organisms can be totally ordered along chromosomes. Given this assumption, the hypothesis we investigate is the following: *the amount of non-linearity observed in ancestral adjacency graphs is correlated to the level of errors made by earlier steps of the pipeline leading to these graphs, in particular to the amount of errors in gene trees.* We are not claiming that these errors are the only possible source of non linearity. Indeed our ancestral adjacency reconstruction method can make mistakes itself; in particular it is a parsimony method, as such unable to cope with convergent or reverse evolution. However we expect that, if we are provided with real species tree, gene families and gene trees and if gene order has evolved without much convergent evolution, then reconstructing ancestral adjacencies should result in few false positive adjacencies and the resulting ancestral adjacency graphs should be close to linear, *i.e.* most ancestral genes should have at most two neighbors.

The idea of a correlation between the extent of non linearity in ancestral adjacency graphs and the distance to an ideal situation was first introduced by Bérard et al. (2012) to compare two sets of gene trees, and has been used in several works to compare gene trees (Boussau et al., 2013; Patterson et al., 2013; Peres and Crollius, 2015; Duchemin et al., 2017; Anselmetti et al., 2018a) or species trees (Anselmetti et al., 2018a). However, there is so far no systematic study providing a proof of principle. In particular the hypothesis that a better linearity implies that the input gene trees are more accurate has never been assessed. This is what we propose to do. We use simulations of species tree, gene trees, gene sequences and gene orders in two situations, one where gene families evolve by speciation, duplication and loss and one where gene families evolve by speciation, duplication, loss and horizontal gene transfer (HGT). We then perturb the gene trees and species tree in order to measure the effect of the introduced errors on the linearity of ancestral adjacency graphs.

In our first set of experiments, we observe a very strong correlation between the amount of noise introduced in the gene trees and, on one hand, the number of inferred structural mutations of genomes, and on the other hand, the non-linearity of the ancestral adjacency graphs. This tends to confirm our hypothesis and suggests that predicted ancestral genomes could be used to assess the quality of phylogenomics data and could thus provide an important signal to correct gene trees. Indeed, while predicted ancestral genomic features, such as gene content, can always be explained by – possibly highly non-parsimonious and unrealistic – evolutionary scenarios, the non linearity of gene order can not be justified in any way. So we provide an additional, original, quality measure. In a second set of experiments, we observe

that with moderately perturbed gene trees, a local error in the species tree correlates with a burst of non linearity precisely in the ancestral genomes close to the erroneous branch. This burst is very localized and could be used to give a hint on erroneous parts of species phylogenies. Finally, in a third set of experiments, we reproduce gene tree construction pipelines starting from sequence data; our results suggest, based on the linearity score, that integrated phylogenomics methods, including gene tree species tree reconciliation, lead to more accurate results than gene tree reconstruction methods based only on multiple alignments.

2 Methods

Our experiments are based on the analysis of simulated data, providing a clear ground truth on the evolution of a set of gene orders. We first describe these simulations, then the analyses performed on the simulated data.

2.1 Simulations

We used Zombi (Davín et al., 2019) to perform simulations. This program constructs artificial species tree, gene trees evolving along this species tree, extant and ancestral gene orders evolving through genome rearrangements, and gene DNA sequences. Genomes evolve by duplication of one or several genes, losses, horizontal gene transfer, and inversions of segments of several genes. Duplications are tandem or not, according to a parameter, and transfers either replace a homologous gene or consists of an insertion at a random place in the genome. Zombi is interesting for our purpose because it mixes gene based events and genome based events. Moreover it is the only available software which is able to take into account extinct or unsampled species when performing HGTs.

The set of parameters is fully available in the supplementary material of this paper. We simulated one species tree with 151 leaves, 26 of them being extant species, the others being extinct or unsampled species. The ancestral gene order at the root of the tree is composed of a single circular chromosome of $1,000$ genes, with no in-paralogs. From there we simulated two datasets:

- Dataset 1: gene families evolved through speciation, gene duplication and gene loss;
- Dataset 2: gene families evolved through a more comprehensive model including speciation, gene duplication, gene loss and horizontal gene transfer.

For each dataset, we obtained from Zombi the true gene tree for each gene family, together with the gene orders of all extant and ancestral species and the DNA sequence of all genes.

2.2 Correlating non-linearity of adjacency graphs and errors in gene trees

In this first experiment, we introduced errors in gene trees and species tree and measured how this impacts a non-linearity score recorded in the adjacency graphs of the ancestral species. An overview of the whole process is depicted on Figure 1.

2.2.1 Introducing errors in gene trees

For each dataset, we introduced various levels of errors in the true gene trees by applying random local perturbations using Nearest Neighbor Interchanges (NNI) on gene tree branches uniformly at random. The level of noise was controlled by the number of NNI performed,

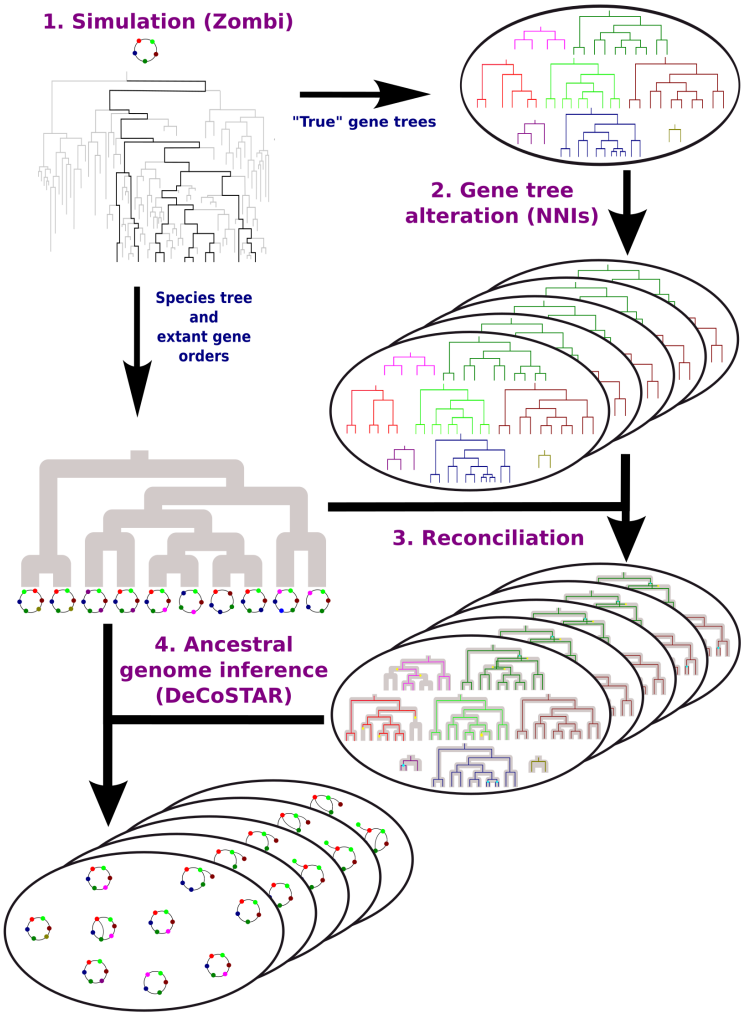


Figure 1 Overview of the simulation/perturbation/reconstruction process and dependencies. We use Zombi to simulate species trees, gene trees, gene orders and gene sequences. We apply some perturbations to gene trees. Then we use DeCoStar to reconcile gene trees (it uses the ecceTERA package) and to construct ancestral adjacencies.

chosen from a Poisson distribution with parameter $\lambda \in \{0.25, 0.5, 1, 2, 3, 5, 7, 10, 20, 30, 50\}$. It follows that we obtained 12 sets of gene trees (the true trees and 11 sets of perturbed trees) for each starting dataset.

For each set of perturbed gene trees we recorded the mean Robinson-Foulds (RF) distance to the true trees. Figure 2 plots the RF distance growing with λ, showing that in this parameter range, there is no saturation of gene tree perturbation, but that the RF distance, more than λ itself, can capture the amount of distortion.

2.2.2 Introducing noise in the species tree

We also looked at the impact of errors in the species tree. To do so, we perturbed the species tree by manually performing a single NNI at an arbitrary branch; we denote by S the true species tree and S_1 the perturbed species tree. Both trees were tested with true and perturbed gene trees.

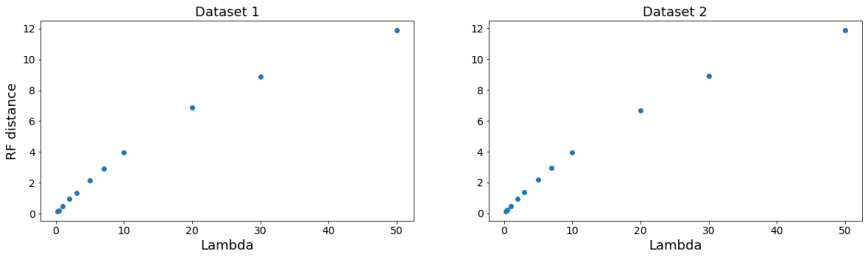


Figure 2 Mean Robinson-Foulds distance as a function of the value of λ.

2.2.3 Reconstructing ancestral gene adjacencies with DeCoStar

DeCoStar (Duchemin et al., 2017) takes as input extant gene orders, gene trees and a species tree. Gene trees are reconciled with the provided species tree using ecceTERA (Jacox et al., 2016), an exact dynamic programming algorithm computing a parsimonious reconciliation. Then ancestral adjacencies are reconstructed for each ancestral species. The principle for this reconstruction is that first extant adjacencies are clustered into families, according to the homology of the corresponding gene extremities. Then for each family of adjacencies, ancestral adjacencies are constructed also with an exact dynamic programming procedure minimizing the cost of gains and breakages of adjacencies; we used default costs for adjacency gain and break (3 and 1).

We added to the program DeCoStar a novel feature, described here for the first time, aimed at handling gene losses without artificially increasing the parsimony cost of an adjacency evolutionary scenario, which is important regarding the linearity score we describe later. This feature consists in iterating the DeCoStar program several times, modifying the costs of creating adjacencies in function of the previous iteration. More precisely, if in the solution computed by the algorithm at iteration i the loss of a gene A, located between two genes B and C, is inferred, then at iteration $i+1$ the gain of an adjacency between genes B and C is free, *i.e.* it does not increase the cost of the evolutionary scenario for the adjacency family containing B and C. This is generalized to any set of consecutive genes located between B and C, being lost concomitantly at iteration i. It can have a significant impact on the linearity score in the case of convergent losses of genes. As we focus on linearity we use this "loss aware" option with two iterations ($i = 2$) in all our experiments.

For each run of DeCoStar, we recorded the number of gene duplications, gene losses and HGTs, as well as the number of gains and breaks of adjacencies.

2.2.4 Non-linearity score

Each run of DeCoStar results in a set of ancestral gene adjacencies. Under the hypothesis that perfect data and a moderate amount of non parsimonious structural evolution will result in linear ancestral genomes, we expect that each gene is the extremity of exactly two adjacencies[2]. In other words it has *degree* 2, where the degree, noted $\deg(g)$ for a gene g, is the number of adjacencies using g as an extremity. Thus we define the *non-linearity score* as the distance from this expectation. For a given ancestral species, the non-linearity score is

[2] This is true for circular chromosomes, and in particular for our current experiments with Zombi. In general chromosomes may be linear and the genes at their extremities are expected to be involved in only one adjacency. However their number is so low compared to a standard gene set that this expectation can be used in practice as well for linear chromosomes

the sum of $|\deg(g) - 2|$ over all vertices g of its ancestral adjacency graph. The non-linearity score for a given experiment is then the sum of the non-linearity scores over all ancestral species.

2.3 Reconstructing gene trees

In our third experiment, for both datasets we reconstructed gene trees for all gene families from the simulated gene sequences, using IQ-TREE (Nguyen et al., 2015) with bootstrap supports on all branches. For Dataset 1 we corrected the IQ-TREE trees with Treerecs (Comte et al., 2020). For Dataset 2 we additionally constructed a sample of gene trees from the sequences using MrBayes (Ronquist et al., 2012) and used the amalgamation option of ecceTERA (Jacox et al., 2016) to construct a single reconciled gene tree from the MrBayes sample, that is, one reconciled gene tree per gene family.

The rationale behind these choices of methodologies is that for gene families evolving under the duplication/loss model (Dataset 1) there are fast methods to obtain reconciled gene trees from IQ-TREE trees, able to correct branch supports, such as Treerecs. For gene families evolving also with HGTs (Dataset 2), where the same problem is NP-complete, we used the amalgamation principle, in a reconciliation framework considering HGTs, which requires to compute a sample of gene trees from a posterior probability.

3 Results

First, we discuss in details our main observation that the non-linearity score is highly correlated with the level of noise introduced in the gene trees. In a second set of experiments, we consider an erroneous species tree and we show that again the non-linearity score increases around the erroneous branch of the species tree, suggesting it could be used to point at species phylogeny errors. Last we consider gene trees reconstructed from sequence data, instead of true gene trees perturbed by random NNIs, and show that reconstruction methods accounting for gene evolution events perform better in terms of non-linearity scores than traditional phylogenetics methods.

3.1 The distance to true trees is highly correlated with the non-linearity score

3.1.1 Overview

Our first result is illustrated in Figure 3: in the two datasets, the non-linearity score grows almost linearly with the mean RF distance between the perturbed gene trees and true trees.

Figure 3 actually represents three scores: the reconciliation score (cost of gene family evolutionary events: gene duplications, gene losses, HGTs), the DeCoStar score (cost of adjacency gains and breakages) and the non-linearity score (see Section 2). We present these three scores, despite the fact that our main interest is in the linearity score, in order to give a broader picture of the impact of noise in the input data on the result of phylogenomic algorithms. In particular, it is interesting to observe that the three scores grow almost linearly with the mean RF between the perturbed gene trees and the true gene trees.

3.1.2 Gene content

An interesting observation is that the reconciliation score grows much faster in Dataset 1 than in Dataset 2. As reconciliation defines the gene content of ancestral species, and it was shown

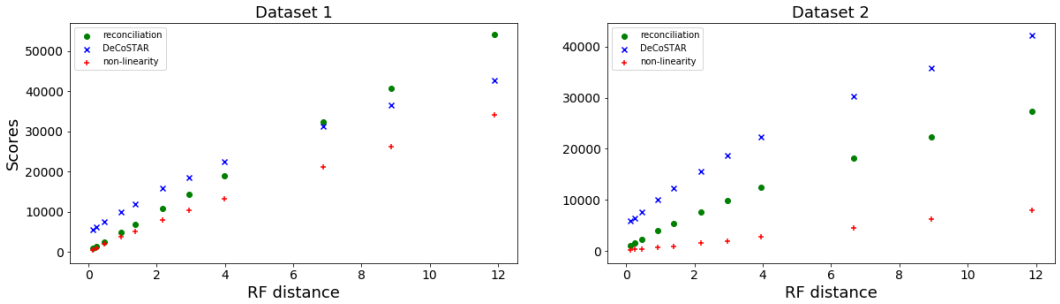


Figure 3 Non-linearity score (red +), DeCoStar score (blue x) and reconciliation score (green circle) as a function of the mean Robinson-Foulds distance between the perturbed gene trees and the true gene trees.

by Hahn (2007) that in a duplication/loss model, errors in gene trees result in an unrealistic gene content of ancestral species, especially for higher nodes of the species phylogeny, we were interested in recording the gene content of ancestral species in both datasets (Figure 4). A somewhat surprising observation is that the patterns of observed gene content deduced from the reconciliations are very different: in Dataset 1, as expected, more ancient ancestral species accumulate genes due to how the parsimonious reconciliation algorithm copes with errors in gene trees, while in Dataset 2, the converse happens, as ancestral species closer to the root of the species phylogeny have less genes, although the variation is less strong than in Dataset 1.

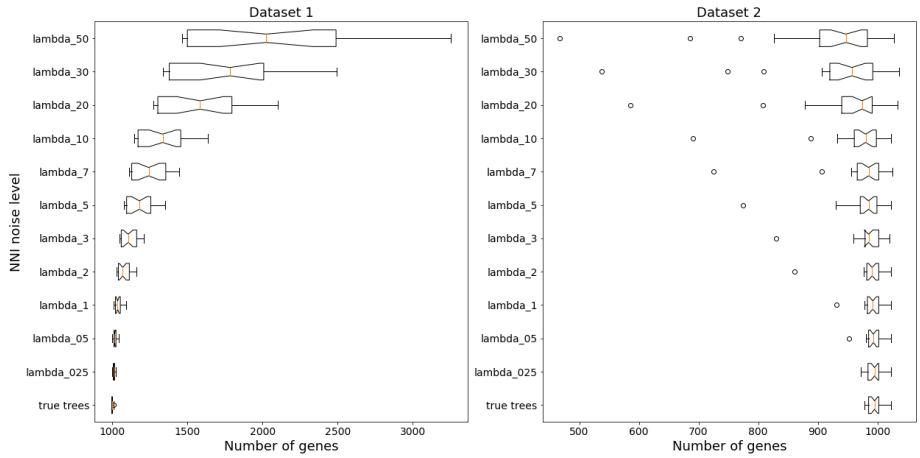


Figure 4 Gene content of ancestral species, in function of the degree of perturbation in gene trees. For Dataset 1 (Left) and Dataset 2 (Right). We see an opposite behavior of DL models (Left) and DTL models (right) with respect to gene content. On one side gene content increases with perturbation, on the other it decreases. In both cases gene content is altered proportionally to the amount of perturbation.

3.1.3 Non-linearity score

We now refine the analysis of the non-linearity score by considering the non-linearity score specific to each internal node of the species tree. We first focus on Dataset 1 and, for the sake

of clarity, we consider only the rates of errors in gene trees of $\lambda \in \{0.25, 0.5, 1\}$ (Figure 5), as they illustrate well the general trend observed for all levels of noise.

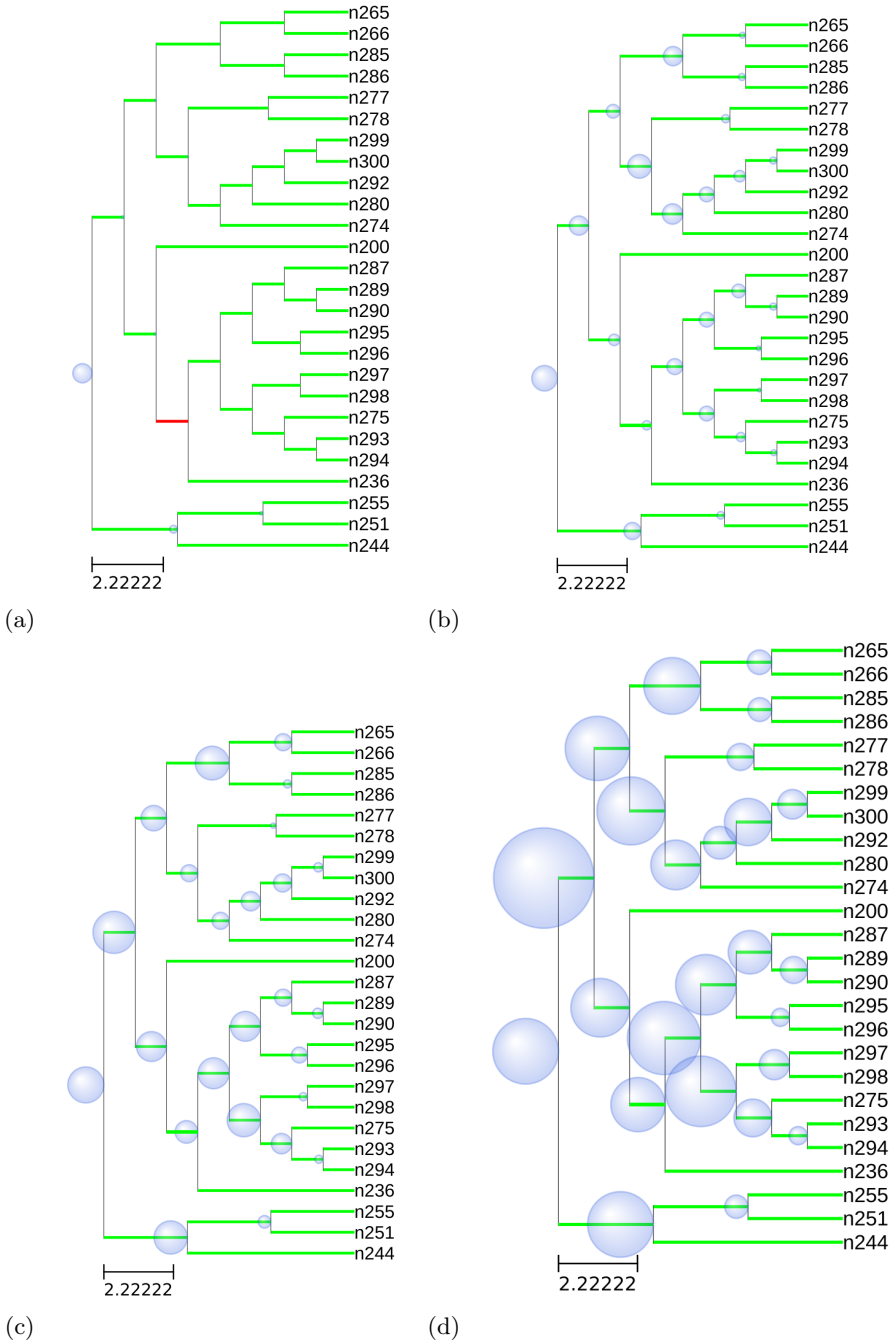


Figure 5 Non-linearity score for Dataset 1 with the true species tree S, with the true gene trees (a), and with λ having value 0.25 (b), 0.5 (c) and 1 (d). The radius of the disks at the internal nodes are proportional to the non-linearity scores. Red branches are the ones that are perturbed (see next section).

The main observation from Figure 5 is that there is a general trend that the non-linearity score increases toward higher nodes of the species tree. It is also interesting to notice that

even with a low level of noise, some lower internal node, such as the roots of cherries (subtrees composed of two leaves) show a non-zero non-linearity score. This suggests that few errors in gene trees are sufficient to create conflicting ancestral adjacencies.

When considering Dataset 2, the effect is rather different, with the root node capturing most of the non-linearity score (Figure 6).

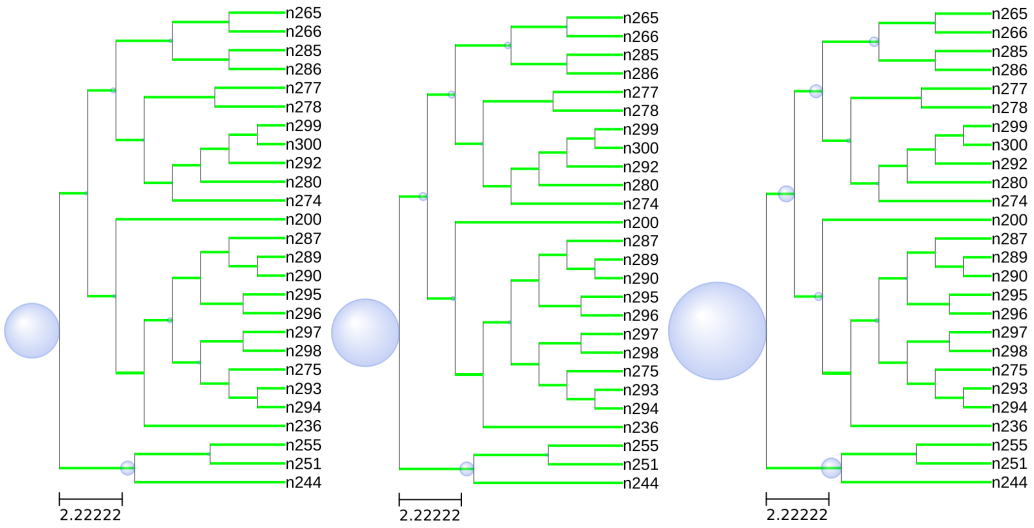


Figure 6 Non-linearity score for Dataset 2, with λ having value 0.25 (Left), 0.5 (Middle) and 1 (Right). The radius of the disks at the internal nodes are proportional to the non-linearity scores. The red branches should be ignored.

Figure 7 below provides another illustration of the difference in terms of non-linearity score variation, as we can observe a much lower magnitude of the score in Dataset 2, as well as a lesser variation compared to Dataset 1.

3.1.4 Discussion

Regarding the interpretation of our observations, an important element is the applicability toward correcting erroneous gene trees. The work described by Hahn (2007) was already a first result toward our hypothesis that scores of phylogenomic algorithms can be correlated to errors in data. Our experiments allow us to go one step further. Indeed, while largely inflated ancestral genomes can be highly unrealistic, one can always consider that there is a non-zero probability that they are correct. A similar remark could apply to the DeCoStar algorithm, that considers individual adjacencies outside of their wider genomic context: adjacency evolutionary scenarios involving a high number of gains and/or breaks could be seen as unrealistic, but not impossible. On the contrary, under the assumptions we outlined in Introduction, a non-zero non-linearity score without false positives ancestral adjacencies is impossible, as genes are linearly or circularly arranged along chromosomes. So if methods are developed with the aim to correct gene trees guided by the reduction of some score, the non-linearity score is a good candidate since its ideal value is known – the closer to zero, the better the gene trees.

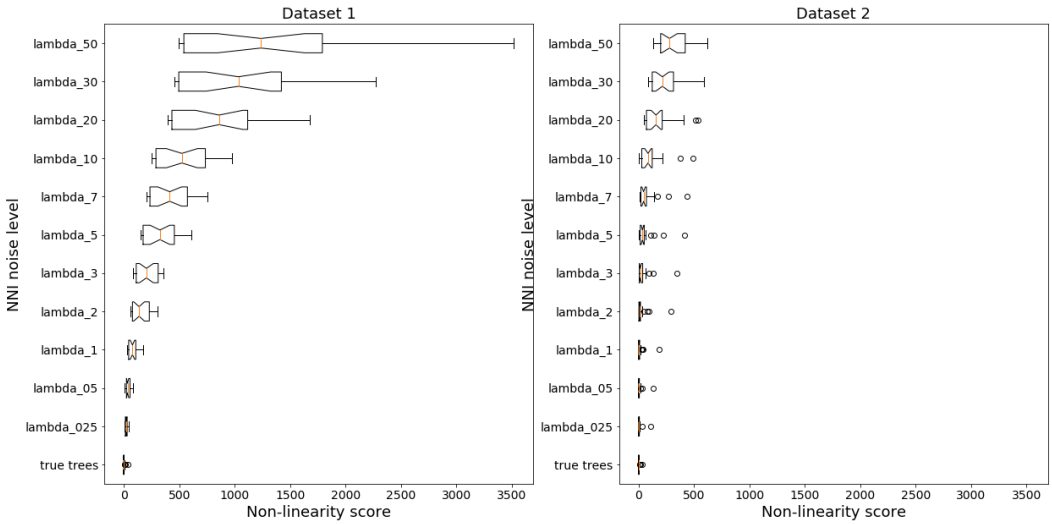


Figure 7 Distribution of the non-linearity score of ancestral species, as a function of the perturbation on gene trees. On the left pannel, for the experiment with only duplications and losses, and on the right pannel, for duplications, transfers and losses. The two pannels have the same scale, in order to illustrate the effect of transfers, in the presence of which the perturbations have a lower effect.

3.2 Non-linearity point at erroneous branches of the species trees

When considering the species tree S_1 differing from the true species tree by a single NNI, the results we obtained in terms of the correlation of the scores of the different steps of our pipeline (reconciliation, DeCoStar, non-linearity) with the level of noise in the gene trees were similar to the ones described above (Figure 8).

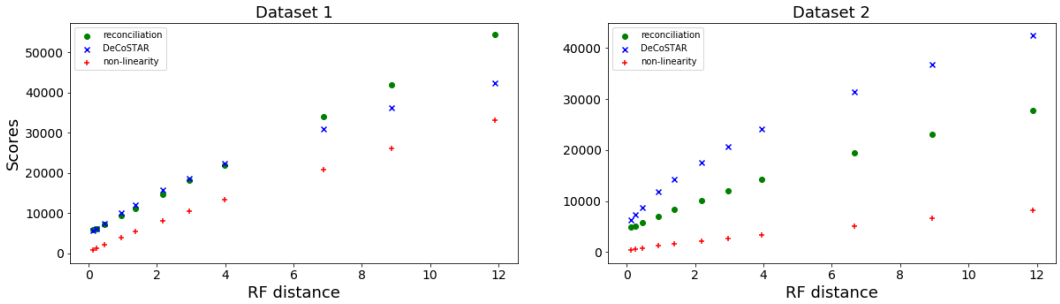


Figure 8 Non-linearity score (red +), DeCoStar score (blue x) and reconciliation score (green circle) as a function of mean Robinson-Foulds distance of perturbed gene trees to true gene trees, with a perturbed species tree.

Moreover, similar to the phenomenon observed when using the true gene trees for Dataset 1, we can observe in Figure 9 that the non-linearity score is greatly inflated around the branch where the NNI was performed, compared to the neighbouring nodes, especially with lower levels of errors in gene trees ($\lambda \in \{0.25, 0.5\}$). This suggests that the non-linearity score can also capture an important signal regarding the accuracy of the species tree.

Next, for Dataset 1 and the true gene trees, we can make two interesting observations by

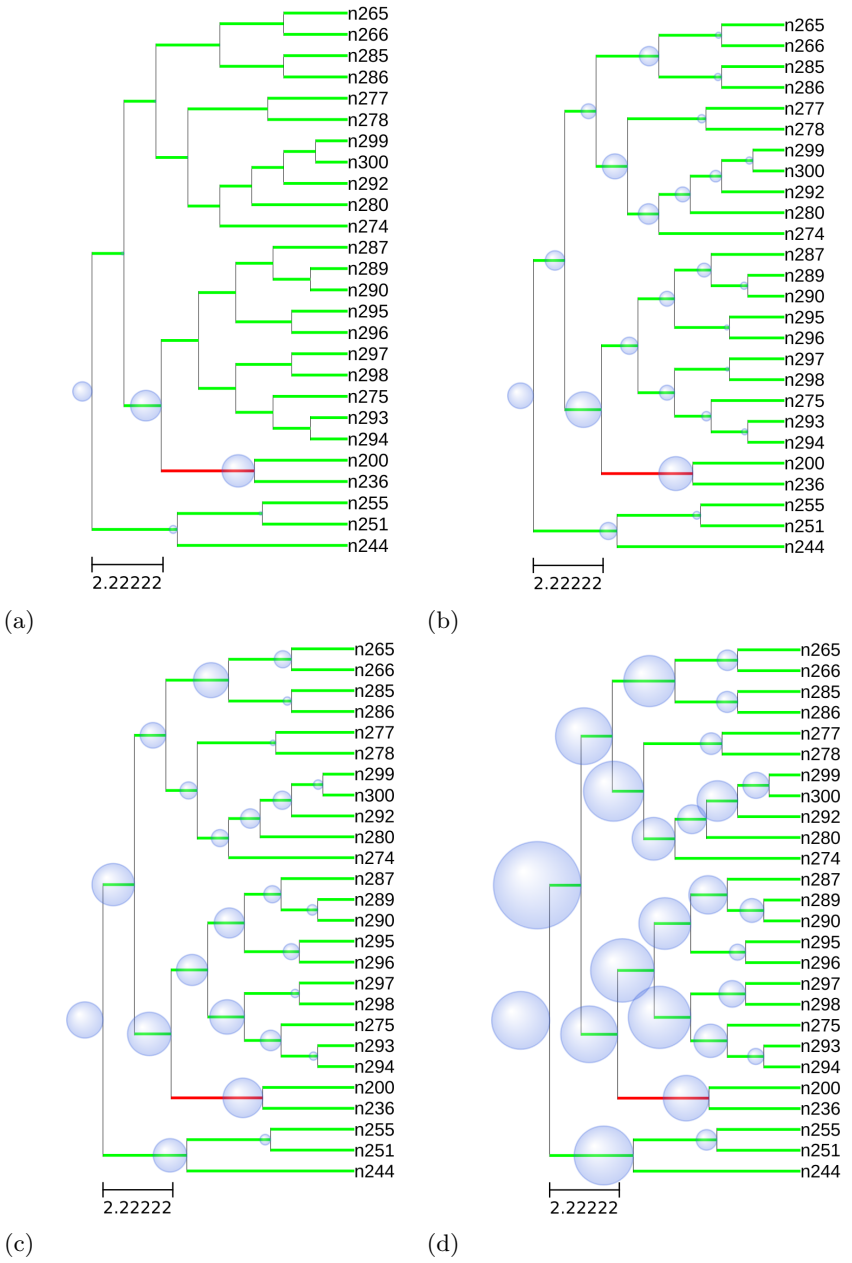


Figure 9 Non-linearity score for Dataset 1 and the species tree S_1, with the true gene trees (a) and with λ having value 0.25 (b), 0.5 (c) and 1 (d). The radius of the disks at the internal nodes are proportional to the non-linearity scores. The branch that has undergone the NNI is shown in red.

comparing the level of non-linearity at each node of the true species tree S, Figure 5(a), and the modified one S_1 (one NNI away), Figure 9(a). First, our assumption that with perfect data from the ground truth (gene families, gene trees, species tree), ancestral adjacency graphs are almost linear, is confirmed. Second, when considering the experiment with the species tree S_1, we can observe a much higher level of conflict, at the branch where the NNI was done, and at its parent.

Last, on Figure 10 we present the same information for the dataset where gene trees

evolved with HGT (Dataset 2). We can observe a similar trend of a level of conflict increasing in higher nodes in the species tree and a strong impact of the NNI performed onto S to obtain S_1, in terms of conflict around the NNI branch.

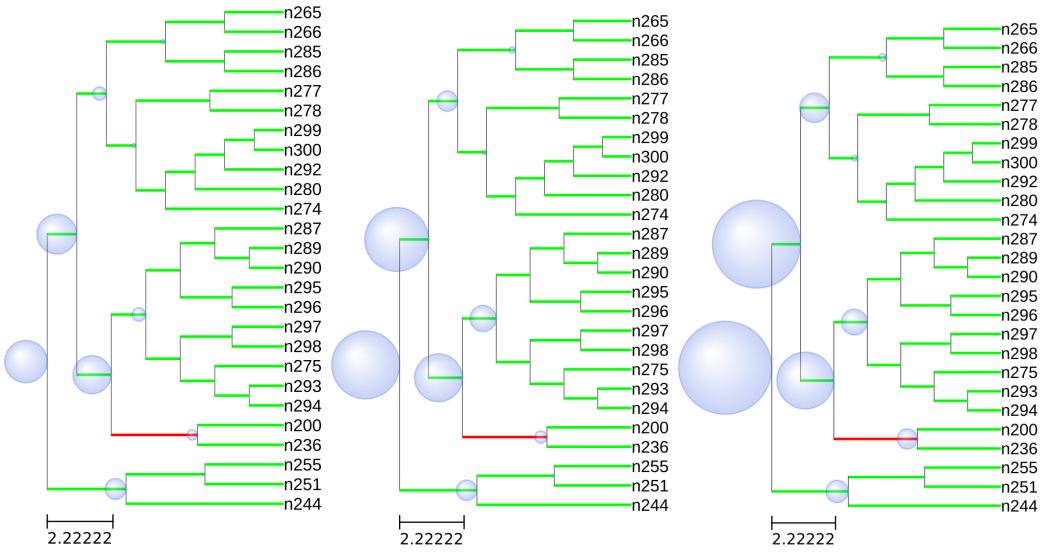


Figure 10 Non-linearity score for Dataset 2 with species tree S_1, with λ having value 0.25 (Left), 0.5 (Middle) and 1 (Right)

Note that these linearity scores are obtained with an error in the species tree uncorrelated with potential errors in gene trees. This is an unrealistic assumption because we can expect that the cause of errors in a species tree are seen as well in the gene trees, or even come from correlated errors in the gene trees. Further tests are needed to control for this effect. Nonetheless some previous observations on *Drosophila* showed that the linearity score was the highest precisely at the most debated node (Semeria et al., 2015), or in *Anopheles* that the linearity score was lower with a species phylogeny agreeing with the gene trees (Anselmetti et al., 2018b). There is an agreeing body of arguments showing that the linearity score is a promising proxy for species phylogeny.

3.3 Phylogenomic methods to reconstruct gene trees

Finally we compared two ways of constructing gene tree sets, in terms of statistics discussed in the previous sections. First we reconstructed gene trees, using IQ-TREE, from multiple sequence alignments of the gene families simulated with Zombi. Next we used integrated phylogenomic methods including the principle of reconciliation to reconstruct gene trees (see Chapter 3.2 [Boussau and Scornavacca 2020]). We refer to Section 2.3 for a description of the methods used with each dataset.

Figure 11 shows the ancestral gene content (number of ancestral genes by species) distribution and the linearity of ancestral genomes, according to different sets of gene trees (true trees, IQ-TREE trees and trees reconstructed with a reconciliation method).

As in our previous experiments, we can observe that the behavior is different in the duplication/loss (Dataset 1) and duplication/loss/HGT (Dataset 2) cases. For Dataset 1, the gene content is unrealistically higher with IQ-TREE trees, confirming the remark by Hahn (2007) discussed above that errors in gene trees could affect the gene content of ancestral

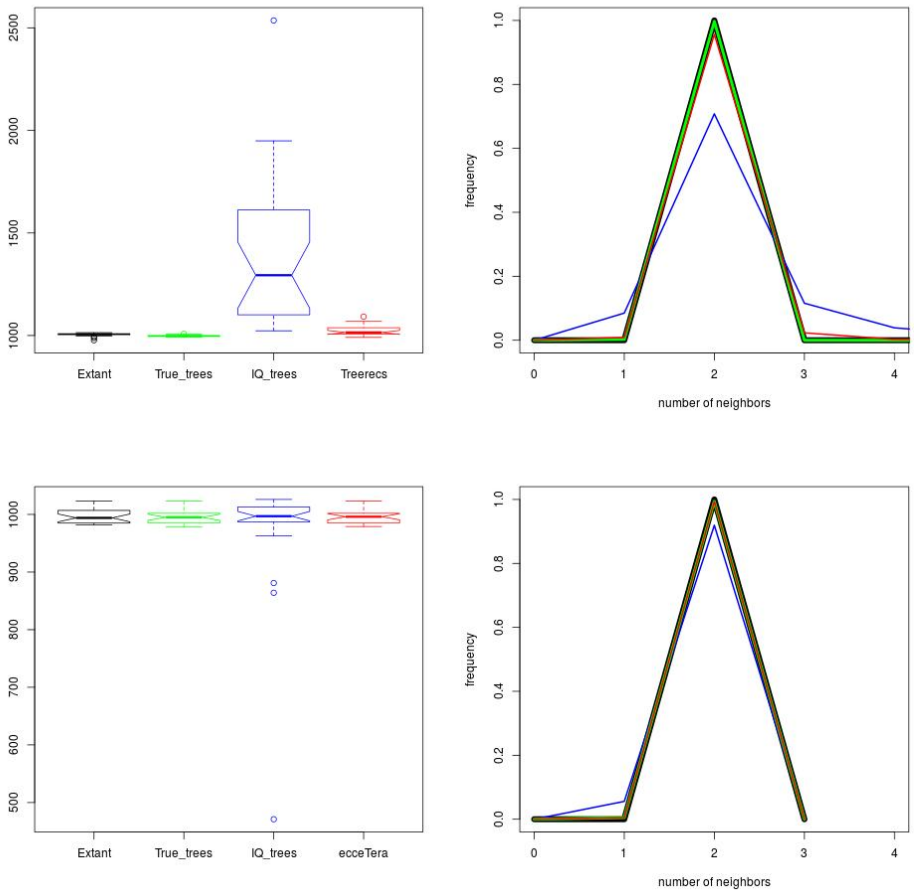


Figure 11 (Left) Distribution of extant (black) and ancestral (other colors) gene contents, computed with true trees (green), IQ-TREE (blue) and Treerecs (red). (Right) Frequency of degrees (number of neighbors) of ancestral genes. (Top) Dataset 1 (with duplications and no transfer), the number of ancestral genes is vastly overestimated and the linearity is lowered if trees are computed from sequences only (IQ-TREE). Both are improved by the reconciliation method. (Bottom) Dataset 2 (with transfers and duplications), the number of ancestral genes is underestimated and the linearity is slightly lowered if trees are computed from sequences only (IQ-TREE). Both are improved by the reconciliation methods.

species. The linearity is almost the same for true trees and reconciled trees, and much worse for IQ-TREE trees. This confirms the intuition present in several former papers (Boussau et al., 2013; Anselmetti et al., 2018b) that the linearity and gene content could serve as an indicator of the quality of gene trees. For Dataset 2 the results are similar but present significant and interesting differences. First, contrary to Dataset 1, gene content is lower for low quality trees, instead of higher. The linearity differences, if present, are much less marked. It seems that the possibility of HGTs in the reconciliation methods can "correct" the errors in gene tree topologies and gives nonetheless almost correct gene numbers and ancestral genome linearity, which, if true, would be an interesting case of robustness of a pipeline to errors in preliminary steps.

4 Conclusion

The present work was motivated by the observation, in previous works from our group, that our efforts to improve species trees or gene tree sets had effects on the linearity of ancestral genomes. We thus formulated the hypothesis that the non-linearity would be a good indicator of the quality of the input data, especially gene trees. In order to explore this hypothesis, we designed a set of experiments on simulated data where the level of noise in the considered trees (gene trees and species tree) is controlled. This allowed us to test our starting hypothesis, and we indeed observe that there is a strong correlation between the non-linearity score and the level of noise. As discussed above, this observation could have practical applications, where non-linear structures in the adjacency graphs of ancestral gene orders could provide starting points to correct gene trees or the species tree.

From a methodological point of view, our general idea can be described as follows. Facing a computationally intractable problem (reconstructing ancestral gene orders in a parsimony framework), one can relax some biological constraints (here the fact that chromosomes are linear or circular gene orders) in order to gain computational tractability; then the inconsistencies observed in the obtained solution with regard to the relaxed biological constraints open a window toward improving the input data. A few examples exist of this kind of serendipitous approaches. For example, one can think to horizontal gene transfers: biology would impose time-consistency on reconstructed transfers, however the problem of inferring time-consistent transfers is NP-hard (Hallett et al., 2004). Finding transfers while allowing them to be time-inconsistent can be solved polynomial in polynomial time (Jacox et al., 2016; Bansal et al., 2018). And, as shown by Chauve et al. (2017), it seems that, similarly to the way we interpret the non-linearity of ancestral gene orders, the level of time inconsistency is correlated with the quality of the input data.

Our work is limited to this proof of principle. We devised experiments only within a small range of parameters, that were chosen to show the possibility of using linearity as diagnosis and its limits. We do not cover all biological conditions. In particular the effect of errors in alignments, gene clustering or annotations have not been investigated, and can be the object of future work.

References

Abrouk, M., Murat, F., Pont, C., Messing, J., Jackson, S., Faraut, T., Tannier, E., Plomion, C., Cooke, R., and Feuillet, C. (2010). Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends in Plant Science*, 15(9):479–487.

Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., and Vialette, S. (2009). On the approximability of comparing genomes with duplicates. *Journal of Graph Algorithms and Applications*, 13(1):19–53.

Anselmetti, Y., Duchemin, W., Tannier, E., and Bérard, S. (2018a). Phylogenetic signal from rearrangements in 18 Anopheles species by joint scaffolding extant and ancestral genomes. *BMC Genomics*, 19(2):96.

Anselmetti, Y., Luhmann, N., Bérard, S., Tannier, E., and Chauve, C. (2018b). Comparative methods for reconstructing ancient genome organization. In Setubal, J. C., Stoye, J., and Stadler, P. F., editors, *Comparative Genomics: Methods and Protocols*, volume 1704 of *Methods in Molecular Biology*, pages 343–362. Springer New York.

Babcock, E. B. and Navashin, M. S. (1930). *The Genus Crepis*, volume 6. Bibliographia Genetica.

Bansal, M. S., Kellis, M., Kordi, M., and Kundu, S. (2018). RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, 34(18):3214–3216.

Bérard, S., Gallien, C., Boussau, B., Szöllősi, G. J., Daubin, V., and Tannier, E. (2012). Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, 28(18):i382–i388.

Blin, G., Chauve, C., Fertin, G., Rizzi, R., and Vialette, S. (2007). Comparing genomes with duplications: A computational complexity point of view. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4):523–534.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Boussau, B., Szöllősi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, 23:323–30.

Chauve, C., Rafiey, A., Davin, A. A., Scornavacca, C., Veber, P., Boussau, B., Szöllősi, G. J., Daubin, V., and Tannier, E. (2017). MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers. *Peer Community In Evolutionary Biology*.

Comte, N., Morel, B., Hasic, D., Guéguen, L., Penel, S., Boussau, B., Daubin, V., Scornavacca, C., Gouy, M., Stamatakis, A., Tannier, E., and Parsons, D. (2020). Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. submitted.

Czech, L., Huerta-Cepas, J., and Stamatakis, A. (2017). A critical review on the use of support values in tree viewers and bioinformatics toolkits. *Molecular Biology and Evolution*, 34(6):1535–1542.

Davín, A. A., Tricou, T., Tannier, E., de Vienne, D. M., and Szöllősi, G. J. (2019). Zombi: A phylogenetic simulator of trees, genomes and sequences that accounts for dead lineages. *Bioinformatics*, 36(4):1286–1288.

Dobzhansky, T. and Sturtevant, A. H. (1938). Inversions in the chromosomes of Drosophila Pseudoobscura. *Genetics*, 23(1):28–64.

Duchemin, W., Anselmetti, Y., Patterson, M., Ponty, Y., Bérard, S., Chauve, C., Scornavacca, C., Daubin, V., and Tannier, E. (2017). Decostar: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. *Genome Biology and Evolution*, 9(5):1312–1319.

Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.

Fuller, Z. L., Koury, S. A., Phadnis, N., and Schaeffer, S. W. (2018). How chromosomal rearrangements shape adaptation and speciation: Case studies in Drosophila pseudoobscuraand its sibling species Drosophila persimilis. *Molecular Ecology*, 28(6):1283–1301.

Groussin, M., Daubin, V., Gouy, M., and Tannier, E. (2016). Ancestral reconstruction: Theory and practice. In *Encyclopedia of Evolutionary Biology*, pages 70–77. Elsevier.

Hahn, M. W. (2007). Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biology*, 8(7):R141.

Hallett, M. T., Lagergren, J., and Tofigh, A. (2004). Simultaneous identification of duplications and lateral transfers. In *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology, 2004, San Diego, California, USA, March 27-31, 2004*, pages 347–356.

Hoff, M., Orf, S., Riehm, B., Darriba, D., and Stamatakis, A. (2016). Does the choice of nucleotide substitution models matter topologically? *BMC Bioinformatics*, 17(1):143.

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., and Scornavacca, C. (2016). eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058.

Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T., and Poon, A. F. Y. (2016). Ancestral reconstruction. *PLOS Computational Biology*, 12(7):1–20.

Kovác, J. (2014). On the complexity of rearrangement problems under the breakpoint distance. *Journal of Computational Biology*, 21(1):1–15.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lemaitre, C., Braga, M. D. V., Gautier, C., Sagot, M.-F., Tannier, E., and Marais, G. A. B. (2009). Footprints of Inversions at Present and Past Pseudoautosomal Boundaries in Human Sex Chromosomes. *Genome Biology and Evolution*, 1:56–66.

Ma, J., Ratan, A., Raney, B. J., Suh, B. B., Zhang, L., Miller, W., and Haussler, D. (2008). DUPCAR: reconstructing contiguous ancestral regions with duplications. *Journal of Computational Biology*, 15(8):1007–1027.

Murat, F., Peer, Y. V. d., and Salse, J. (2012). Decoding Plant and Animal Genome Plasticity from Differential Paleo-Evolutionary Patterns and Processes. *Genome Biology and Evolution*, 4(9):917–928.

Necsulea, A. (2020). Phylogenomics and genome annotation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.1, pages 4.1:1–4.1:26. No commercial publisher | Authors open access book.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Patterson, M., Szöllosi, G. J., Daubin, V., and Tannier, E. (2013). Lateral gene transfer, rearrangement, reconciliation. *BMC Bioinformatics*, 14(S-15):S4.

Peres, A. and Crollius, H. R. (2015). Improving duplicated nodes position in vertebrate gene trees. *BMC Bioinformatics*, 16(3):A9.

Philippe, H., de Vienne, D. M., Ranwez, V., Roure, B., Baurain, D., and Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, 283.

Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C., and Salse, J. (2019). Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biology*, 20(1):29.

Rajaraman, A. and Ma, J. (2016). Reconstructing ancestral gene orders with duplications guided by synteny level genome reconstruction. *BMC Bioinformatics*, 17(S-14):201–212.

Ranwez, V. and Chantret, N. (2020). Strengths and limits of multiple sequence alignment and filtering methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.2, pages 2.2:1–2.2:36. No commercial publisher | Authors open access book.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.

Semeria, M., Tannier, E., and Guéguen, L. (2015). Probabilistic modeling of the evolution of gene synteny within reconciled phylogenies. *BMC Bioinformatics*, 16(Suppl 14):S5.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Sturtevant, A. H. (1921). A case of rearrangement of genes in Drosophila. *Proceedings of the National Academy of Sciences U S A*, 7(8):235–237.

Tannier, E., Zheng, C., and Sankoff, D. (2009). Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, 10.

Timoshevskiy, V. A., Severson, D. W., deBruyn, B. S., Black, W. C., Sharakhov, I. V., and Sharakhova, M. V. (2013). An integrated linkage, chromosome, and genome map for the yellow fever mosquito aedes aegypti. *PLOS Neglected Tropical Diseases*, 7(2):1–11.

Yang, Z. and Zhu, T. (2018). The good, the bad, and the ugly: Bayesian model selection produces spurious posterior probabilities for phylogenetic trees. arXiv preprint https://arxiv.org/abs/1810.05398.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761.

Zhou, L., Lin, Y., Feng, B., Zhao, J., and Tang, J. (2017). Phylogeny analysis from gene-order data with massive duplications. *BMC Genomics*, 18(7):760.

Zhou, X., Shen, X.-X., Hittinger, C. T., and Rokas, A. (2018). Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Molecular Biology and Evolution*, 35(2):486–503.

Chapter 3.1 The Sources of Phylogenetic Conflicts

Dominik Schrempf
Department of Biological Physics, Eötvös Loránd University
Pázmány P. stny. 1A, H-1117 Budapest, Hungary
dominik.schrempf@gmail.com
http://orcid.org/0000-0001-8865-9237

Gergely Szöllősi
MTA-ELTE "Lendület" Evolutionary Genomics Research Group
Department of Biological Physics, Eötvös Loránd University
Pázmány P. stny. 1A, H-1117 Budapest, Hungary
Evolutionary Systems Research Group, Centre for Ecological Research
Hungarian Academy of Sciences, 8237 Tihany, Hungary
ssolo@elte.hu
http://orcid.org/0000-0002-8556-845X

Abstract

Recombination breaks up the evolutionary history between genomic regions and, as a result, the evolutionary history of different genomic regions may differ. In fact, conflicting phylogenetic signal between genes is commonplace. The reasons for conflicting signal may be statistical or systematic in nature. In order to avoid strongly supported but incorrect inferences driven by systematic error, use of appropriate phylogenetic methods accounting for these processes is of fundamental importance.

This chapter reviews possible causes of phylogenetic conflict between genes. Processes generating conflict, including gene duplication and loss, horizontal gene transfer, hybridization, and incomplete lineage sorting are presented using classic examples. In particular, we discuss compelling evidence for whole genome duplications in fish, as well as plants, and the role of horizontal transfer in the spread of antibiotic resistance. Finally, building on the material presented, we show how these processes lead to phylogenetic conflict, and how they can be described by phylogenetic models.

How to cite: Dominik Schrempf and Gergely Szöllősi (2020). The Sources of Phylogenetic Conflicts. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 3.1, pp. 3.1:1–3.1:23. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Funding DS and GJSz received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under grant agreement no. 714774 and the grant GINOP-2.3.2.-15-2016- 00057.

1 Introduction

The evolution of present-day life, that is, the causal events of the diversity of observed sequences, has been a complicated and intertwined process. Consequently, resolving the events and their order from observed hereditary sequences is challenging. This section sets out to describe some of these evolutionary processes such as transmission of genes from parents to offspring during reproduction, gene duplication and loss or horizontal gene

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 3.1; pp. 3.1:1–3.1:23

A book completely handled by researchers.
No publisher has been paid.

transfer, and continues to corroborate that they affect phylogenetic inference. In fact, if these processes are ignored, they can lead to false inferences with strong statistical support (see Chapter 2.1 [Simion et al. 2020]).

Before moving on, we have to reflect on what phylogenetic conflicts actually are. The concept of *homology* (Greek; *homos*, same; *logos*, relation) plays a pivotal role in answering this question. Evolutionary characters are defined to be either homologous or analogous (Greek; *ana*, up to, upwards), if they have shared or no shared ancestry, respectively. The definition of homology appears simple but harbours several layers of hidden complexity. For example, it is possible that ultimately all characters have a single ancestor. The decision about whether two sequences are homologous or not involves a statistical test where the null hypothesis is that the observed similarity simply arose by chance.

Consequently, not single DNA or amino acid characters are compared, but sequences of characters. For this reason we focus on the larger, abstract phylogenetic unit of a *gene*. We think of a gene as a set of characters, mostly but not necessarily adjacent along the sequence, which is inseparably transmitted together from parent to offspring, or, more abstract, from donor to recipient. Recombination reshuffles the genetic sequence during the production of offspring, and thereby separates different genomic fragments so that their evolutionary histories are partially independent. We assume that recombination acts only between and not within genes. The assumption that genes are passed on as a whole is reasonable because recombination within genes is likely to break function, and so, is likely to be deleterious.

Usually, pair-wise distances between sequences are determined using BLAST (Altschul et al., 1997). The resulting distance matrix can be interpreted as a completely connected graph. The vertices are the characters, and the edge weights correspond to BLAST scores. Typically, a threshold on edge weights is applied to produce a sparser graph without edge weights. Then, graph clustering algorithms and other criteria such as sequence overlap are used to identify sub-graphs representing tightly connected homologous clusters (Miele et al., 2011; Li et al., 2003; Enright et al., 2002). Admittedly, the threshold discriminating between sequences being part of the same or of different homology clusters depends on user-defined parameters.

The outcome of homology search are distinct sets of homologous genes which are termed *gene families*. Genes are sorted into different families if we have no means of finding common ancestry, either because the genes are not similar enough, or because intermediate homologous genes have not been sampled. After all, it is possible that all genes belong to a single gene family. Gene families are a treasure trove of information to reconstruct ancestry but they bear some enigmas. Given that we have decided that two genes belong to the same gene family, we can track their histories back in time until they merge. At the merging point, different events may have happened. First, and probably most frequently, the corresponding event may be associated with a cell division during reproduction. In this case, the ancestor and the two descendants are all members of the same species, and the two descending genes are at the same genomic position, which we call *locus*. Loosely, a locus is the genomic location where a gene resides (Gillespie, 2004). As a result of speciation, the two descendant genes may end up in two different species (Section 2.4). We call genes related by coalescence or speciation *orthologs* (Greek; *orthos*, in a straight line, true, correct, see Chapter 2.4 [Fernández et al. 2020]). Second, the merging point may correspond to some form of gene duplication (Section 2.2). The two descending genes co-exist at different loci in the genome of the same individual. Genes related by duplication are termed *paralogs* (Greek; *para*, besides). Whole genome duplication is a special, concerted form of gene duplication; genes related by whole genome duplication are termed *ohnologs* (in honor of

Susumu Ohno). Third, the merging point may correspond to some form of horizontal gene transfer (Section 2.2). The ancestral and one of the descendant genes are of the same species, which is termed *donor*. The other descendant gene is part of the genome of an individual belonging to a different species, the *recipient*. In this case, we speak of *xenologs* (Greek; *xenos*, foreign). Finally, genes brought together by inter-species hybridization (Seciont 2.3) are called *homeologs* (Greek; *homeo*, similar).

Biological mechanisms of duplication and transfer will be discussed in the next section. The evolutionary history, or phylogenetic relationships, of all genes within a gene family can be depicted in a *gene tree*. We use the term tree to refer to a tree-like object including topology and branch lengths. Similar to gene trees, relationships of species are depicted in a *species tree*. Increasingly sophisticated models of sequence evolution (see Chapter 1.1 [Pupko and Mayrose 2020]) have allowed accurate reconstruction of gene and species trees.

Importantly — and this brings us back to our original problem of defining phylogenetic conflict — recombination breaks up the histories of different genes within a gene family and, all the more, between gene families. Within a few generations linkage is erased and genes evolve independently of each other. Hence, phylogenetic methods usually assume free recombination between genes (see Chapter 3.4 [Bryant and Hahn 2020]). In the light of recombination, the presence of paralogs, ohnologs, xenologs, and homeologs — that is, gene duplication together with loss, horizontal gene transfer, and hybridization — can cause disagreement between a gene tree and the species tree. As a matter of fact, it can even cause disagreement about the species tree within a single gene tree, for example, when histories of two ancient paralogs are different. Disagreement of a gene tree with the species tree or disagreement within a gene tree about the species tree is the definition of *phylogenetic conflict*. However, we usually have no *a priori* knowledge about the type of homology, and so do not know if two genes are orthologs, paralogs, ohnologs, xenologs, or homeologs.

Further, genes within a gene family have different, independent coalescent times. Consequently, orthologous genes not coalescing before the previous speciation event can lead to phylogenetic conflict (see Chapter 3.3 [Rannala et al. 2020]). Moreover, free recombination between genes also implies that gene families have independent histories, aside from evolving along the same species tree. It follows that phylogenetic conflict will also be manifested as disagreement between the gene trees themselves. As a matter of fact, phylogenetic conflict is mostly observed as disagreement between gene trees because the species tree is unknown, as it cannot be reconstructed directly.

For example, the cumulative effects of gene duplication, gene loss and horizontal gene transfer result in severe conflict between six gene trees obtained from protein coding sequences (Philippe and Forterre, 1999). Further, all gene trees conflict with the famous three domain tree previously obtained from ribosomal RNA by Woese et al. (1990). Even within Eukaryotes, gene trees reconstructed from different protein sequences exhibit conflicts, and the order of emergence of basal groups depends on the rate of evolution of the protein used.

It seems that we are confronted with an unsolvable conundrum at the heart of which are recombination, gene duplication and loss, horizontal gene transfer, and hybridization. In this chapter we aim to elucidate how independent coalescence, gene duplication and loss, horizontal gene transfer, and hybridization can cause phylogenetic conflict. Given that we understand the causes and processes responsible for phylogenetic conflict, it is our task to develop methods that can assess the likelihood of gene trees supporting a given species tree (see Chapter 3.2 [Boussau and Scornavacca 2020]). In the end, most phylogeneticists are interested in resolving the best-supported species tree, and elucidate the origin and evolution of life.

2 Processes causing phylogenetic conflict

Evolutionary change of genetic sequences is introduced by a series of mutations which may spread in a population, either by chance or because they provide selective advantages. Mutations affecting more than one DNA base are called structural mutations. The size of structural mutations ranges from a few bases to billions of bases. Structural mutations which remove parts from the genome and add parts to the genome are called deletions and insertions, respectively. Many biological processes can cause the creation of new gene structures within the genome of an individual (for a review, see Long et al., 2003). Some of these processes create perfect to near-perfect gene copies, others may only copy parts of genes or combine/extend different genes (e.g., exon-shuffling). Further processes include retroposition, where a gene is inverted by reverse transcription with subsequent insertion back into the genome of the individual, insertion of mobile elements into genes, or fusion and fission of genes.

Even though the creation of new gene structures involves complex biological processes, their quantification in mathematically tractable models requires significant simplification. To our knowledge, most phylogenetic models only consider perfect copying of genes, either in the same individual (gene duplication), or copy/cut and paste into a coexisting individuals from a different species (horizontal gene transfer). In accordance with current phylogenetic models, we distinguish in the following between (1) gene origination, which is the formation of a novel genetic sequence establishing a new gene family and (2) gene birth and death, which are the addition of a gene to a gene family and the removal of a gene from a gene family, respectively. Further, inter-species hybridization and incomplete lineage sorting are discussed.

2.1 Gene origination

Across all three domains of life, up to 30 percent of the genes within an organism have no known homologs, and are presumably of recent origin (Tautz and Domazet-Lošo, 2011; Dehal, 2001). First, such *orphan genes* can emerge through the creation of new genetic material for example by gene duplications due to errors during recombination, the acquisition of extrinsic genes, or by transposition mechanisms. The generation of new genetic material is followed by a phase of fast adaptive evolution leading to divergence beyond the threshold of homology searches.

Second, gene families may originate due to *de novo* evolution. Random sequences from non-coding regions may form cryptic functional sites that could subsequently come under regulatory control. In this case, the creation of de novo genes may be as simple as having a mutation at a single base that activates transcription of a downstream stretch of DNA. The activated sequence may fortuitously code for a protein enhancing fitness. Initially, *de novo* evolution was deemed highly unlikely, but several cases with compelling evidence have been observed and reported. For example, several human specific genes have been detected which correspond to non-transcribed regions in other primates (Knowles and McLysaght, 2009). The question is: "Is more likely that these genes have been switched on in humans, than switched off in all other primates"? The patterns of the observed population data provide evidence that the open reading frames are functional.

2.2 Gene birth and death

Changes in gene copy number in a gene family are a frequent mutational event (Reams and Roth, 2015). However, the exact mechanisms are difficult to discover, because they vary with genomic position. Further, determination of gene birth rates, that is, the number of events happening per unit time, is hard. Furthermore, the mechanisms of gene birth and death are distinct, and therefore, different rates can arise. Conceptually, gene birth and gene death denote the increase and decrease of gene copy number, respectively. Below, we give a brief summary of gene birth through gene duplications, horizontal gene transfer, and hybridization, and gene death which is synonymously called gene loss.

Gene duplication

Gene duplication (Ohno, 1970), which is the emergence of a heritable copy of a gene within a genome of an individual, is an integral constituent of the evolution of biological complexity. Gene duplication is prevalent across the tree of life and has greatly shaped the hereditary material of present-day organisms (Kondrashov et al., 2002). In fact, gene duplication is the most common source of new genes in Eukaryotes. The presence of a gene copy can have detrimental as well as beneficial effects. Often, gene duplications are present in some but not all individuals of natural populations resulting in a situation called copy number variation. Although we have fundamental knowledge about gene duplications on the functional and the genomic level (Conant and Wolfe, 2008), knowledge about their emergence, and maintenance is incomplete. In the following, common biological mechanisms leading to gene duplication are outlined.

1. Unequal crossing over is an event where the breaks during recombination happen at different positions on the chromosomes which are then misaligned during meiosis.
2. During replication slippage, the DNA polymerase erroneously changes position and copies a part of the chromosome twice.
3. Retrotransposition is a process where messenger RNA is reverse transcribed to DNA which is integrated back into the genome.
4. Repetition of protein domains is a pattern observed within many genes. Repeated domains can be caused by exon-shuffling, which is the duplication of exons as a result of recombination between non-homologous sequences.
5. Transposition of active mobile elements (Long et al., 2003; Lynch, 2007) may also lead to novel genes partially containing duplicated genetic material.
6. The nucleus of polyploid organisms contains more than two sets of chromosomes. Whole genome duplication is an extreme type of mutation where the gamete produced during meiosis carries the entire diploid genome rather than the haploid one. Whole genome duplication results in so-called tetraploidy, a condition where each chromosome is present four times in the nucleus. Repeated whole genome duplication leads to octaploidy, although copies might be lost between the consecutive whole genome duplications resulting in a different number of chromosome copies. Whole genome duplication is more common in plants than animals, but see the discussion below about fish and jawed vertebrates.
7. In a similar manner, the fusion of two genomes during hybridization of two species results in an allo(tetra)polyploid and is discussed in Section 2.3.

In prokaryotes, the process of gene duplication is less well understood, but several methods allowing assessment of duplication rate are now available (Reams and Roth, 2015). Often, the methods involve beneficial multi-step processes.

Most newly created genes will be removed by genetic drift within a few generations. Only a very small fraction of novel genetic material will rise in frequency and eventually become fixed in a population. In the case of gene duplications, the two gene copies may be redundant and consequently under relaxed selection. As a result, one gene copy may again be lost by a deletion or the accumulation of loss-of-function mutations (see below). More interesting, one gene copy may evolve a novel biological function; a process termed neofunctionalization. The two gene copies may also subfunctionalize, and perform separate functions which, in concert, fulfill all original functions and may provide more (see Chapter 4.2 [Robinson-Rechavi 2020]).

Whole genome duplications have been a major confounding factor in phylogenetic analyses (Van de Peer et al., 2017). For example, two rounds of whole genome duplications (2R) in the last common ancestor of all jawed vertebrates were the cause of a fourfold increase of vertebrate genes (Sidow, 1996). The 2R hypothesis was intensely debated, but is now widely accepted (Meyer and Van de Peer, 2005). Even more, a third round of whole genome duplication has happened in fish (Meyer and Schartl, 1999; Brunet et al., 2006), but evidence was not always conclusive. For example, Robinson-Rechavi et al. (2001) analyze the evolutionary history of 35 gene families in fish. Seven gene families follow a pattern consistent with an ancestral whole genome duplication, whereas eleven gene families show duplications that had most likely happened after the divergence of the considered fish species, and 19 gene families do not show any signs of duplications.

Whole genome duplications are even more common in plants. For example, there is evidence that the genome of the common ancestor of angiosperms was duplicated three times. In fact, this may be the reason for the morphological and ecological diversification of angiosperms. Polyploidy increases biological complexity and the amount of genetic material subject to natural selection. The resulting phenotypic variation, mostly caused by overall differences in expression levels between diploid and polyploid individuals may lead to selective advantages in polyploids. For example, polyploidy has been repeatedly discussed in the context of major evolutionary transitions, and hybrid vigor. Further, it has been hypothesized that polyploidy is a major driver of adaptive radiation of species (De Bodt et al., 2005).

Inference of ancient whole genome duplications is difficult due to saturation of synonymous distance (Tiley et al., 2018). Recently, probabilistic methods have been developed to infer whole genome duplications (Zwaenepoel and Van de Peer, 2019) employing amalgamated likelihood estimation (ALE, Szöllősi et al. 2013a; Chapter 3.2 [Boussau and Scornavacca 2020]).

ALE is based on conditional clade probabilities, which roughly correspond to the observed frequency distribution of clades. These probabilities can be calculated from a collection of gene trees, or from trees yielded during a bootstrap analysis, or during an MCMC analysis. Importantly, gene tree uncertainty is accounted for, and also unobserved gene tree topologies can be evaluated.

Horizontal gene transfer

In contrast to vertical inheritance of genes from parents to offspring, organisms can also incorporate foreign genes, or variant copies of foreign genes from distant relatives through a process termed horizontal gene transfer. A successful horizontal gene transfer event requires the genetic material to be successfully released by the donor, transported to the recipient, acquired and incorporated into the genome of the recipient, and expressed in a way that benefits the recipient. Several processes enable the required succession of hereditary events.

1. Transduction (Latin; *trans* - across, beyond; *duco*, to lead, to conduct) is the import of viral DNA through agents such as bacteriophages, probably even as a result of infection.
2. Conjugation by plasmids (Lating; *con*, together; *jugum*, yoke; "yoke together") is horizontal transfer involving direct cell-to-cell contact, for example, through surface appendages with transfer of plasmids.
3. Transformation (Latin; *forma*, to shape, to form, to direct) is the uptake of free, extracellular DNA.
4. Gene transfer agents are bacteriophage-like elements integrated in the donor genome (Gogarten and Townsend, 2005; Stewart, 2013; Soucy et al., 2015). Gene transfer agents are sometimes under regulatory control by the donor. They package random DNA fragments from the donor and transport them to a recipient. Horizontal transfer by gene transfer agents differs from transduction in that, unlike bacteriophages, gene transfer agents are unable to transport all required genes to reproduce themselves.

Knowledge about the different forms of transfer is important because the ranges, and, consequently, their signatures differ.

Most interestingly, horizontal gene transfer played a key role in the identification of DNA as the molecular basis of inheritance. The famous experiment by Griffith (1928) showed that a non-virulent bacterial strain can incorporate genetic material of a heat-killed virulent strain, and subsequently cause disease. Avery et al. (1944) identified DNA to be a substance transferring genetic information by transformation. When antibiotic resistance spread unexpectedly fast across many different enteric strains (Davies and Davies, 2010; Davies, 1996), it was widely appreciated that horizontal gene transfer can not only be induced in the lab, but is of general importance in the evolution of bacterial genomes.

Direct observation of horizontal gene transfer happens rarely, and so, evidence of its occurrence needs to be collected independently from the traces that are left behind in the molecular sequences themselves. Of course, we expect that a horizontally transferred gene exhibits high resemblance between donor and recipient, and that the gene will be limited to the descendants of the initial donor and recipient. Especially if donor and recipient are distantly related, unduly high levels of resemblance between restricted sub-groups of otherwise unrelated species should capture the attention of methods detecting horizontal gene transfer.

Early studies collecting evidence for horizontal gene transfer analyzed the nucleotide compositions, and patterns of codon usage bias (Ochman et al., 2000). Genes with sequence characteristics departing significantly from the rest of the considered genome were classified as recent transfers. The detected amount of transferred DNA varied greatly between virtually 0 to nearly 17 percent in the 19 bacterial and archaeal genomes analyzed. This result is most likely an underestimation, because transfers from species with similar sequence characteristics cannot be detected. In fact, there is growing evidence, that horizontal gene transfer has played a major evolutionary role and has integrally shaped bacterial and archaeal genomes, as well as their diversification and speciation patterns. In fact, a significant proportion of bacterial and archaeal genetic diversity has been acquired through horizontal gene transfer (Abby et al., 2012; Lerat et al., 2005). The high frequency of horizontal gene transfer among prokaryotes can lead to phylogenetic relationships that are more net-like than tree-like. The importance of horizontal transfer in eukaryotes was still a topic of dispute until recently, when suspected examples of horizontally transferred genes were detected in metazoans including sponges, cnidarians, rotifiers, nematodes, molluscs, arthropods (Boto, 2014), and even humans (Crisp et al., 2015). Additionally, significant amount of transfer was observed across the kingdom of fungi (Szöllősi et al., 2015). For reviews, refer to Daubin

and Szöllősi (2016) or Husnik and McCutcheon (2017).

Finally, a transfer without recognizable homologs may be interpreted as gene origination. Possible reasons can be, for example, rapid divergence following the transfer event, gene loss in the donor species, erroneous homology search, or incomplete sampling and sequence availability.

Gene loss

Gene loss is the removal of existing genes from a gene family. On the one hand, gene loss can be a sudden mutational event caused, for example, by unequal crossing over during meiosis, or transposition of mobile elements. On the other hand, gene loss can be a slow process. Nonsense mutations creating truncated proteins or frameshifts, as well as missense mutations affecting crucial amino acid positions lead to the initial inactivation of a gene. The so-called *pseudogenization* is followed by a progression of deletion events with small fitness effect. Non-functional genes, may they be the result of inactivation, or non-functional duplicates are called pseudogenes. The number of pseudogenes can be large. For example, the human genome has nearly as many pseudogenes as functional genes (Lynch, 2007).

Gene loss greatly influences the gene content of genomes and contributes to the divergence of related species, next to other processes such as mutation. The potentially caused phenotypic diversity indicates that gene loss may be an adaptive evolutionary change (Albalat and Cañestro, 2016). Especially in bacteria and archaea, genome size is a strong fitness determining factor (*less-is-more* hypothesis). Likewise, unexpected high levels of intra-species variation of gene losses have been observed.

2.3 Hybridization

Occasionally, species evolve from hybrid crosses between two different ancestral species, an event termed inter-species hybridization. Inter-species hybridization is especially common in plants. The resulting genomic and phenotypic features reveal the two ancestral sources. Species histories exhibiting inter-species hybridization have rejoining branches, and as such, are not tree-like but correspond to more general networks called directed acyclic graphs. The respective gene trees can take various topologies depending on which gene copies are kept or lost.

Bread wheat is one of our most important staple crops and has been cultivated for more than ten thousand years. The evolution of bread wheat comprises several hybridization events (Pont et al., 2019). The genome of bread wheat consists of three closely related subgenomes, which are usually denoted (AABBDD). A first hybridization between wild *Triticum urartu* (AA) and a species of the *Aegilops speltoides* lineage (BB) produced a allotetraploid species, of which Durum wheat is a direct descendant. Subsequently, a second hybridization event with the wild *Aegilops tauschii* species (DD) formed the present day allohexaploid genome.

Further, analysis of transcriptome data revealed pervasive ancient hybridization event in relatives of bread wheat (Glémin et al., 2019). Detection of hybridization events was performed using a hybridization index (Meng and Kubatko, 2009). The hybridization index measures the likelihood that a hybridization event, and not incomplete lineage sorting, is the cause of the observed phylogenetic conflicts. The analysis revealed that the overlooked wild species *Aegilops mutica* was involved in the first hybridization event leading to an allotetraploid (AABB) which further developed into bread wheat.

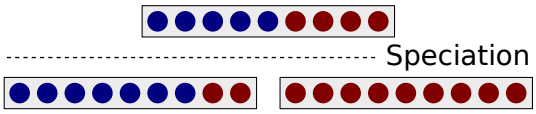


Figure 1 Incomplete sorting of alleles during a speciation event of a population of 9 individuals (circles). The blue and the red allele coexist at a specific locus in the population during the speciation and are incompletely sorted. The left daughter species contains two individuals with the red allele. At the considered locus, these two individuals are more closely related to the individuals in the right species than to the individuals containing the blue allele in their own species.

An ancient interspecies hybridization was also detected in the baker's yeast lineage (Marcet-Houben and Gabaldón, 2015). The authors propose that the resulting hybrid was forced to undergo a subsequent whole genome duplication to regain fertility.

Another form of hybridization is reassortment with viruses. Host cells simultaneously infected by two virus strains may assemble new viral particles whose origins are mixed. Some genetic material may originate from the first strain, other genetic material from the second.

2.4 Incomplete lineage sorting

In this section, we will describe a process fundamentally different from gene duplication and loss, horizontal gene transfer, or hybridization because it operates on the population level, and leads to conflict between orthologous genes. To facilitate the following exposition, we introduce the term allele which is a variant of a genetic sequence at a specific locus. Here, we mean different orthologous genes of the same species, but in a different context, alleles can also be different nucleotides, amino acids, or even whole chromosomes. The word allele is a short form of *allelomorph* (Gree; *allelo*, mutual, each other; *morph*, form) and emphasizes that we can discriminate between one or the other variant of an entity at the same locus in the genome of the considered individuals.

Different alleles can coexist in a population potentially for a long period of time spanning speciation events. Conceptually, a binary speciation sorts the alleles at a specific locus into the first or the second daughter species. Alleles with more variants fully participating in the speciation, maybe because they are partly causing the speciation, are completely sorted and no allele is present in both daughter species. In contrast, and because of recombination, coexisting alleles can be incompletely sorted during a speciation event such that they are present in both daughter species (Figure 1). For example, the individuals containing the red allele in the first daughter species are more closely related to individuals in the second daughter species than to individuals in their own species carrying the blue allele. Of course, because of recombination, the misleading affinity is only observed at this specific locus. When analyzing more loci, individuals within a species will be closer related to each other than to individuals of the other species (see Chapter 3.4 [Bryant and Hahn 2020]).

The process described above is referred to as *incomplete lineage sorting*. Incompletely sorted alleles coexisting over multiple consecutive speciation events may lead to phylogenetic conflict in that the corresponding gene tree supports a different topology than the species tree. For example, Scally et al. (2012) estimated that up to 30 percent of the genome of humans, chimpanzees, and gorillas has higher support for either one or the other of the two wrong species trees, and not for the correct species tree. That is, if we turn the argument around, only 70 percent of the genome of humans, chimpanzees, and gorillas supports the correct species tree.

3.1:10 The sources of Phylogenetic Conflicts

What are the factors determining the prevalence of incomplete lineage sorting? First, recombination is a prerequisite. Further, at a specific locus, incomplete sorting of alleles is likely when alleles coexist in a population, that is, when more alleles than a single one have higher frequency. A well-known measurement of genetic variation is the heterozygosity H of a locus, which is the probability of sampling two different alleles (e.g., Gillespie, 2004). For a haploid population with neutral variation, the heterozygosity is reduced per generation by *genetic drift*

$$\Delta H_D \approx -\frac{1}{N} H, \tag{1}$$

where N is the size of the considered population, and increased per generation by *mutations* happening with rate u per locus and generation.

$$\Delta H_M \approx +2u(1-H). \tag{2}$$

Consequently, incomplete lineage sorting is prevalent if the population size and the mutation rate are large. Only then is sufficient variation generated compared to the rate at which it is removed by genetic drift. Moreover, incompletely sorted alleles need to coexist over multiple speciation events, in order to cause disagreement between a gene tree with the species tree. In this case, it is not enough that the population size is large, but also that the number of generations between speciation events is low. In particular, shorter internal branches of the species tree when measured in number of generations indicate higher prevalence of incomplete lineage sorting.

The name incomplete lineage sorting originates from the term *lineage* which denotes the line of descent from a common ancestor. The concept of a lineage is especially important when viewing the process backwards in time. Then, incomplete lineage sorting is manifested by lineages of alleles in the same species which do not coalesce within that species but only in an ancestor. For this reason, the term *deep coalescence* is often used to describe incomplete lineage sorting. Deep coalescence only leads to phylogenetic conflict, if the coalescent event involves a lineage ultimately leading to a different species.

3 Phylogenetic description

Development of appropriate models is key to understanding observations and collecting evidence for hypotheses. First, we remind ourselves that phylogenetic conflict is only an issue if genes are not co-transmitted from parents to offspring. For example, mitochondrial genomes are passed without recombination simplifying phylogenetic analysis. However, exclusive analysis of the mitochondrium is unsatisfactory, because the mitochondrium only contains a limited amount of genes and statistic errors are to be expected. In contrast, the eukaryotic nuclear genome contains a vast amount of informative sites but is continuously broken up by recombination.

Similarly, data sets including bacteria and archaea cannot be analyzed, because they also show recombination as a result of horizontal gene transfer combined with homologous recombination. The resulting presence of genes not being orthologs is a pervasive problem (e.g., Page, 2000). One approach to analyze data including genes not being co-transmitted is to concatenate putative orthologs and hope the corresponding phylogenetic signal outweighs the one of spuriously utilized paralogs, xenologs, and so on. Nevertheless, much data is ignored with this procedure. Full exploitation of data including recombining genes requires methods appropriate for analysis of paralogs (e.g., Page, 2000), ohnologs, xenologs, and homeologs (e.g., Szöllősi et al., 2012).

3.1 Homologous group sizes

Historically, search for homologous genes was only performed within species. In this section, the term *homologous group* will be used to denote a set of homologous genes within a species. We refrain from using the term gene family because gene families usually have members in several species. Early methods sought to describe the frequency distribution of the homologous group sizes which follows a power law characterized by long tails (Huynen and van Nimwegen, 1998). In particular, homologous groups with ten or more genes are more abundant than what would be expected from analyzing the frequency of homologous groups of low to moderate size (Szöllősi and Daubin, 2012). Remarkably, the distribution is very similar across bacteria, archaea, and eukaryotes, indicating shared universal features of the underlying processes responsible for the creation and removal of genetic material.

We can improve our understanding about the origins of the observed power law in the frequency distribution of homologous group sizes using stochastic linear birth and death processes (Yule, 1925).

▶ **Definition 1.** A linear birth and death process is a stochastic process that describes the evolution of an integer state variable. The considered system is a number $N_t \in \{0, 1, 2, \ldots\}$ of units at time t evolving according to the following rules (Kendall, 1949; Bartholomay, 1958).

1. The sub-units generated by a unit develop in complete independence of each other.
2. A unit existing at time t multiplies by binary fission with probability $\lambda dt + o(dt)$, and dies with probability $\mu dt + o(dt)$ in the following time-interval of length dt.
3. All units in the population exhibit the same *birth rate* λ and the same *death rate* μ.

In detail, a birth event adds one unit to the population, and a death event removes one unit from the population. Waiting times until the next birth or death event are independently and identically distributed with exponential distributions with predefined birth and death rates λ and μ, respectively. The birth and death rates are shared across all units and independent of the number of units N_t. Linear birth and death processes are particular in that the waiting time until the next event necessarily becomes smaller the more units are present in the population.

In the context of our discussion, the units of the birth and death process are genes within a homologous group, birth events correspond to gene duplication or horizontal gene transfers events, and death events correspond to gene loss events. Of course, as described above, the instantaneous creation of identical gene copies in birth and death processes is only a rough approximation of biological gene birth and death. First, death rates are usually estimated to be higher than birth rates. Second, and more important, mathematical analysis shows that we cannot explain the observed power law with linear birth and death processes alone. That is, there is no set of birth and death parameters that can possibly explain the long tails of the frequency distribution of homologous group sizes. Rather, we need to relax the assumption of independence and employ general birth and death processes where the rates may depend on the total number of units in the population. We can only explain the long tails of the observed frequency distribution of homologous group sizes when letting the death rate, which is usually larger than the birth rate, decrease with the homologous group size so that it approaches the birth rate for large homologous groups (Karev et al., 2002).

The power law can also be obtained as the stationary distribution of a birth and death process with origination of homologous groups (Reed and Hughes, 2004). Specifically, this model is a superposition of two layers of stochastic processes. First, a stochastic process similar to a pure birth process, which has a death rate of zero, describes the origination of

homologous groups. Thereby, homologous groups randomly duplicate at a given rate. Naturally, the number of homologous genes may differ between homologous groups. Hence, this process involves non-identical units and is not a classical, birth and death process. Second, for each homologous group, a separate linear birth and death process with parameters λ and μ describes the number of genes in the respective homologous group. The stationary distribution of this two-layered process exhibits a stretched exponential if $\lambda \leq \mu$, and a power law if $\lambda > \mu$. A stationary distribution exists, because homologous groups are removed when the last gene dies. In conclusion, we fail to obtain the observed power law when treating genes within homologous groups independently. However, the frequency distribution of homologous group sizes can be described by relatively simple stochastic processes such as general birth and death processes, or using a two layered stochastic process modeling origination of homologous groups. Notice that both models treat different homologous groups independently.

At the same time, it is evident that the evolutionary histories of homologous groups of the same species but different gene families are correlated because of common decent along a shared species tree. Further, homologous groups of the same gene family and of closely related species are expected to have more similar sizes than homologous groups of the same gene family but from distantly related species. In this respect, a *gain-duplication-loss* model has been developed which uses a fixed species tree inferred from the concatenated alignments to assess the likelihood of homologous group sizes within gene families (Csűrös and Miklós, 2009; Csűrös, 2010). In the gain-duplication-loss model, gain is the pendant to horizontal gene transfer, but because only gene counts are considered, and not the gene trees themselves (see below), the origin of transferred genes is unknown. and all we can observe, is a gain in gene count. Ignoring the gene tree, may lead to biases in inference of gene transfer (Szöllősi et al., 2015). Rate variation can be accounted for with a discrete gamma distribution similar to the treatment with substitution models. Also, branch-wise gain, duplication and loss rates can be used. An application to archaea and bacteria shows that birth and death rates are similar across the two analyzed domains of life and that death rates seem to be larger than birth rates (Szöllősi and Daubin, 2012).

3.2 Multi-sequence alignments

More recently, advances in sequencing technologies and improvement of clustering and alignment algorithms used in homology search have led to the identification of numerous gene families with available sequence data. The analysis of the hereditary sequences themselves is much more intriguing than the analysis of their mere quantity. The sequence data is usually prepared in form of multi-sequence alignments (Figure 2) which contain a vast amount of information. In general, there is one multi-sequence alignment per gene family, and each sequence is labeled with the corresponding species and a label of their own, since there can be more genes per species. We are not concerned with species delimitation (Rannala and Yang 2013; Yang and Rannala 2010; Chapter 5.5 [Rannala and Yang 2020) in this chapter. Typically, the alignment does not contain information about the homology relationships of the genes, and thus, about the loci. Here, we use loci in form of integers solely to encode the type of homology, and not relative position. That is, locus 1 is not necessarily left of or close to locus 2. However, genes at the same locus are assumed to be orthologous, and genes at different loci are assumed to be paralogous, ohnologs, xenologous, or homeologs. Further, common multi-sequence alignments of gene families only provide information from a reference genome and not about the genetic variation of genes. In detail, they do not contain sequences from different individuals. As a side note, the term *genome* has been used

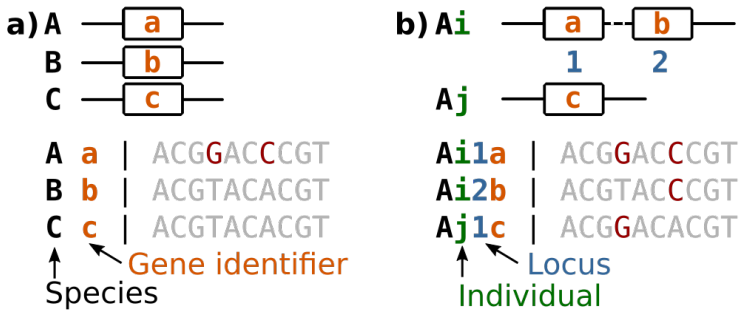


Figure 2 Toy observation of a gene family with corresponding nucleotide multi-sequence alignment containing three genes of length ten. The horizontal lines represent genomes of sampled individuals and rectangles symbolize genes. The dashed line indicates that the genes may not be close to each other (in fact, they may be on different chromosomes). The genomes of the individuals and the genes are labeled with their corresponding species (capital black letter). The genes additionally have labels (orange lowercase letter). Nucleotide variants are emphasized in red. a) The alignment contains no information about the homology relationship of the genes. b) Here, species A contains three genes from two individuals, (green letter), as well as positional information about the locus (blue number). Usually, gene family alignments neither contain population data nor positional information.

in ambivalent ways. First, genome can refer to the complete genetic material present in an individual. Second, and in a more abstract way, genome can refer to the genetic material of a species, a set of species, or more generally, a set of individuals; for example, "the genome of humans" or "archaeal genomes". Even though we are aware of this homonym, we were not able to completely avoid it. However, we aim to be explicit when describing the sources of phylogenetic conflicts in this section. That is why we focus on individuals themselves, for example, we examine "genes present in an individual" (and not in a genome).

For a given multi-sequence alignment of a gene family, we seek to coherently describe possible evolutionary histories. This task involves the construction of a gene tree, and most importantly, the event types at the nodes of the tree. We have already briefly discussed some types of events that can happen at gene tree nodes and which are accounted for by current evolutionary methods. For example, we can classify gene tree nodes as gene duplications or horizontal gene transfers, and add unobserved nodes corresponding to gene losses so that the observed gene tree agrees with the species tree; this process is called gene tree species tree reconciliation. Most gene tree, species tree reconciliation approaches involve two steps: (1) the reconstruction of gene trees with maximum likelihood methods, and (2) the reconciliation thereof using maximum parsimony methods minimizing the total number of gene duplications and losses (see Chapter 3.2 [Boussau and Scornavacca 2020]).

We will now present details on how gene duplication and loss, horizontal gene transfer, and hybridization affect the genes and genomes under consideration (Figure 3). Cell division during reproduction, or coalescence when viewed backwards in time, is the creation of two new individuals of the same species both of which contain the original gene. The locus of the genes remains unchanged. Gene duplication adds a copy of the same gene into the one produced daughter individual. The novel gene copy is inserted at a new locus. The genetic sequences of the copies are identical, and so we do not know which of the genes is to be found at the new locus. We may have to take this combinatorial fact into account during model design. Whole genome duplication corresponds to a concerted, massive gene duplication but can otherwise be described in the same way as simple gene duplication.

3.1:14 The sources of Phylogenetic Conflicts

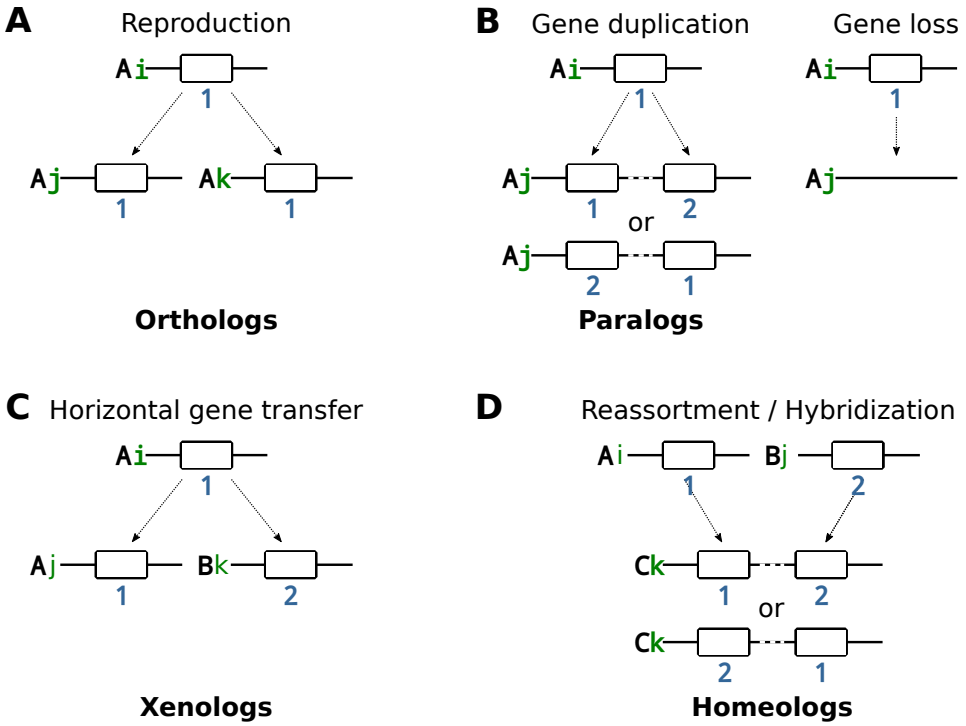


Figure 3 Depiction of how phylogenetic methods model cell divison during reproduction, or coalescence when moving backwards in time, gene duplication and loss, and horizontal gene transfer. The horizontal lines represent genomes of considered individuals and the rectangles symbolize genes. The dashed line indicates may not be close to each other. The sequences of the genes are unaffected by the events. The individuals have a label (green letter) and are assigned to a species (black capital letter). The labels of the genes have been left blank. A hypothetical locus number is written below the genes in blue.

Horizontal gene transfer inserts a gene copy into a coexisting but foreign genome of a different species at a new locus. It has been argued that a horizontally transferred gene can take over the function of a previously existing gene in the recipient. In this case, the absence of purifying selection on the preexisting gene copy can induce divergence or loss of the preexisting gene copy. The result is an event called replacement transfer. There are models that exclusively allow replacement transfers because of their biological relevance and because a replacement transfer corresponds to a specific topological move called subtree pruning and regrafting (Hasić and Tannier, 2019).

Similar to the remark about whole genome duplications above, inter-speices hybridization can be modeled like a concerted, massive horizontal gene transfer event combining two ancestral species. Whole genome duplications, and inter-species hybridization events are visible not only on the gene tree, but also on the species tree.

Finally, gene loss simply removes a gene from the genome. Gene loss is not directly observable in gene trees that we reconstruct, because branches leading to loss events are pruned from the tree. However, gene loss is an important constituent of phylogenetic models because, as we will see below, it can lead to phylogenetic conflict either together with the other discussed processes, or on its own.

A consequence of the considerations above is the following thought: If we had information about the loci of the genes in a given multi-sequence alignment, we could greatly reduce

the number of possible evolutionary histories explaining the data. Orthologs must be from different genomes but the same locus, gene duplication and loss involves genes from the same genome but different loci, and horizontal gene transfer involves genes from different species and possibly different loci. Note that turning this argument around, probabilistic models accounting for the events discussed above are informative about the homology relationships of the genes in the given multi-sequence alignment. Next to elucidating the gene tree, the detection of orthologs, paralogs, and xenologs is an important application (see Chapter 3.2 [Boussau and Scornavacca 2020]). Further, we only know of one model describing the actual synteny of loci, that is, the physical co-localization of loci (Delabre et al., 2018). Thereby, gene duplication and loss can affect segments spanning more than one locus.

4 Summary and discussion

In summary, the term gene loosely denotes a stretch of hereditary sequence passed on as a whole. Two genes with detectable shared ancestry are homologous to each other. A set of homologous genes is called a gene family. A gene family usually spans many species and can have more than one gene per species. The phylogenetic history of a gene family can be depicted in a gene tree. The lineage of a gene is the path from the gene, to the root of the gene tree. The type of homology of two considered genes is defined by the event happening when the lineages of the two genes join in the past. The types of homology we have discussed are (1) orthologous genes related by cell division during reproduction, (2) paralogous genes related by gene duplication, (3) ohnologous genes related by whole genome duplication, (4) xenologous genes related by horizontal gene transfer, and (5), homeologous genes related by inter-species hybridization.

Phylogenetic conflict is the complete or partial disagreement of a gene tree with the species tree. The species tree is usually not known, and so, phylogenetic conflict is either observed as disagreement about the species tree within a single gene tree, or disagreement between different gene trees. All discussed homology relations can cause phylogenetic conflict. For example, reproduction combined with recombination and mutation can lead to incomplete lineage sorting which is manifested by deep coalescing lineages maybe suggesting a misleading topology.

A prerequisite of conflict between gene trees is recombination. In contrast, co-transmitted genes will not show conflict. For this reason, genomic architecture such as chromosomes, plasmids or nuclear vs cytoplasmic compartments is an important factor. Doubts about exclusive usage of genes as phylogenetic units have been raised (Springer and Gatesy, 2018). For example, smaller units such as exons could be used. Further, knowledge about recombination patterns can help in discriminating between phylogenetic reconstruction errors and truly different gene trees (Reddy et al., 2017). As previously mentioned, conflicting histories between mitochondrial genes are unexpected.

The correct description of homology relationships in phylogenetic analyses is imperative, yet, the abundance of actual phylogenetic conflict is a matter of dispute. The importance of incomplete lineage sorting has been a topic of dispute for many years. After all, Scally et al. (2012) estimated that only 70 percent of the genomes of humans, chimpanzees and gorillas follow the correct species tree. High levels of incomplete lineage sorting are also reported in birds (Jarvis et al., 2014). It has been argued that phylogenetic conflict caused by incomplete lineage sorting may be wrongly estimated by the assumption that genes are passed on as a whole and by disregarding exon and intron structure (Springer and Gatesy, 2018; Mendes et al., 2019). In mammals, phylogenetic conflict caused by incomplete lineage sorting is minor

when observing whole genes but conflicting histories are more pronounced when observing exons in a separate way (Scornavacca and Galtier, 2017). Further, it was shown that for species tree aware methods, the number of inferred gene duplications and horizontal gene transfers depends strongly on the used species tree (Szöllősi et al., 2013a). Although this is expected, caution is warranted when interpreting inferences involving phylogenetic conflict. In general, identification of systematic error as well as statistical error is difficult. For instance, it was postulated that phylogenetic conflict between orthologous mitochondrial genes may mostly be caused by statistical or systematic error (Richards et al., 2018).

Another issue is the relative importance of incomplete lineage sorting, gene duplication and loss, and horizontal gene transfer in causing phylogenetic conflict. The probability of incomplete lineage sorting is high when the number of generations between consecutive speciation events is low. If the average branch length measured in number of generations decreases from the root of the species tree towards the present, incomplete lineage sorting is more prevalent close to the present. If we assume that the species tree evolves according to a pure birth model (Yule tree, Yule, 1925), this assumption is not met since the average branch length on the tree is independent of the position of the branch on the tree (Stadler and Steel, 2012). There are no analytical solutions for the distribution of branch lengths for trees originating from a linear birth and death process (e.g., see Paradis, 2016). However, there is evidence from simulations that internal branch lengths increase compared to terminal branch lengths when increasing the death rate from zero towards values closer to the birth rate. This effect is more pronounced the closer the death rate is to the birth rate. Zhaxybayeva and Gogarten (2004) assume that the tree of life evolved according to a coalescent model. The coalescent model assumes that the total population size is constant and that the time to the next coalescence (moving back in time towards the root) is exponentially distributed with parameter $\binom{n}{2}$, where n is the number of sampled species. As a result, the average branch length increases towards the root. In this case, as well as for the birth and death process, the relative importance of incomplete lineage sorting decreases from the leaves to root of the species tree. Of course, we can only hypothesize about the distribution of branch lengths since we do not know the tree of life.

Spurious phylogenetic conflict can also arise if the reconstruction method suffers from systematic error (see Chapter 2.1 [Simion et al. 2020]). For example, across-site or across-gene composition heterogeneity (Lartillot and Philippe, 2004) can cause topological errors. In general, saturation of sequence distance can lead to long branch attraction artifacts (Felsenstein, 1978). Furthermore, decisions made during homology search may induce spurious phylogenetic conflict and greatly influence the identification of gene losses. In particular, if gene origination is not correctly detected, the lack of genes pertaining to the new gene family in neighboring species not affected by the gene origination, may be incorrectly attributed to massive genes loss. Similarly, undetected gene copies can be misinterpreted as gene losses (Page, 2000).

Another aspect related to this topic is the fact that in all phylogenetic analyses, most species remain unsampled. For this reason, trees originating from birth and death processes including the probability of incomplete sampling at the leaves of the tree have been analyzed (Stadler, 2009). In essence, the sampling probability corresponds to a transformation of the birth and death rates assuming complete sampling. Interestingly, a relatively simple calculation shows that we should expect the donors of most horizontally transferred genes in a considered data set to be members of either extinct or unsampled species (Szöllősi et al., 2013b).

The last example displays a big advantage of using birth and death processes in phylo-

genetic inference. The mathematical study of birth and death processes has a long tradition (Yule, 1925; Kendall, 1949; Bartholomay, 1958; Thompson, 1975; Gernhard, 2008; Stadler and Steel, 2019), and many properties have been derived analytically. For example, the expected number of units with time and the corresponding variance are known. Further, the probability density of birth events and the distribution of the time of origin of reconstructed trees have been derived (Gernhard, 2008). The knowledge can be summarized strikingly in so-called lineage through time plots, which show the average number of lineages of a tree evolving under the birth and death process with time. Further, the probability of death and the probability of no change within a given time can be used to calculate, for example, the probability of a gene tree evolving under the birth and death process within a constraining species tree. However, it is still difficult to simulate aforementioned gene trees conditioned on a specific number of genes per species and possibly also on the time of origin of the process. We can employ a forward process combined with rejection sampling, but computation times are immense for larger trees. The problem becomes even more difficult when accounting for horizontal gene transfer. The main reason is that the assumption of independence of the birth and death process is not met anymore. That is, the sub-units created by a birth event do not evolve in complete independence.

In any case, the combined treatment of incomplete lineage sorting, gene duplication and loss, and horizontal gene transfer is demanding but the unification of phylogenetic and population genetic models promises to improve both our understanding of, and our ability to reconstruct, the tree of life. In general, genetic change traverses three stages, (1) origin in a genome of an individual through mutation, (2) fixation in the respective population, and (3) maintenance in the population. For example, the dynamics of gene duplications differ slightly, in that the fates of the copies are tightly linked and determined by changes accumulated during, or after the fixation phase. Innan and Kondrashov (2010) argue that in order to understand these processes we have to examine the genetic variation of gene copies on the population level. Phylogenetic models, combining the description of gene duplication and loss, and horizontal gene transfer with the description of the evolution of variation in populations such as the multi-species coalescent model could play an important role in this respect. Especially, since they will allow more precise identification of the type of homology, which is an important piece of information that could, for instance, help resolve the complete species tree of animals (Pett et al., 2019).

In this respect, the three tree model by Rasmussen and Kellis (2012) has been seen as a considerable methodological advance (Du and Nakhleh, 2018). Additionally, it is now possible to infer hybridization events (Du et al., 2019; Elworth et al., 2019). Hybridization is when individuals of already diverged species successfully have offspring. The hybrids may be founders of a new, separate species transforming the species tree into a phylogenetic network. Hybridization can also be interpreted as a massive horizontal gene transfer which affects large parts of the genome. On the other hand, we have introduced a different three tree model that is more consistent but assumes no recombination between homologous genes on the same haplotype. Both three tree models use ultra-metric, time-like trees but rate modifiers can be used to account for the different molecular clocks. Also, the three tree models do not account for gene origination on the species tree. This may be problematic with respect to gene families limited to a few species. For example, consider two genes from the same homologous gene family that are present in Human and Elephant, but absent in all other mammals. Is this not already phylogenetic conflict caused solely by gene loss (or horizontal gene transfer)? Another type of event not discussed in this chapter is allopolyploidization. Allopolyploidization is the retention of both parental genomes in the offspring, an event that

is especially likely in plant species.

Phylogenetic trees and networks are the basis of a wide range of outstanding probabilistic methods. Even so, they can be accompanied by methods inspired by the machine learning community. For example, a K-means clustering algorithm was applied to data simulated using the three tree model of Rasmussen and Kellis (2012) to create a classifier that can identify orthologous regions (Knowles et al., 2018). Note that the classifier is not an independent method because it is trained on data simulated using the discussed probabilistic methods. Altogether, it is natural to expect that different gene families tell different stories about the species tree. It will be exciting to see how the treatment of the three most important causes of phylogenetic conflict can be combined in a successful and conclusive way.

References

Abby, S. S., Tannier, E., Gouy, M., and Daubin, V. (2012). Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Sciences*, 109(13):4962–4967.

Albalat, R. and Cañestro, C. (2016). Evolution by gene loss. *Nature Reviews Genetics*, 17(7):379–391.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of Experimental Medicine*, 79(2):137–158.

Bartholomay, A. F. (1958). On the linear birth and death processes of biology as markoff chains. *The Bulletin of Mathematical Biophysics*, 20(2):97–118.

Boto, L. (2014). Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proceedings of the Royal Society B: Biological Sciences*, 281(1777):20132450.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution*, 23(9):1808–1816.

Bryant, D. and Hahn, M. W. (2020). The concatenation question. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.4, pages 3.4:1–3.4:23. No commercial publisher | Authors open access book.

Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950.

Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A., and Micklem, G. (2015). Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology*, 16(1).

Csűrös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912.

Csűrös, M. and Miklós, I. (2009). Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. *Molecular Biology and Evolution*, 26(9):2087–2095.

Daubin, V. and Szöllősi, G. J. (2016). Horizontal gene transfer and the history of life. *Cold Spring Harbor perspectives in biology*, 8(4):a018036.

Davies, J. (1996). Origins and evolution of antibiotic resistance. *Microbiologia (Madrid, Spain)*, 12(1):9–16.

Davies, J. and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews*, 74(3):417–433.

De Bodt, S., Maere, S., and Van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*, 20(11):591–597.

Dehal, P. (2001). Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. *Science*, 293(5527):104–111.

Delabre, M., El-Mabrouk, N., Huber, K. T., Lafond, M., Moulton, V., Noutahi, E., and Castellanos, M. S. (2018). Reconstructing the history of syntenies through super-reconciliation. *Lecture Notes in Computer Science*, pages 179–195.

Du, P. and Nakhleh, L. (2018). Species tree and reconciliation estimation under a duplication-loss-coalescence model. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB 18*.

Du, P., Ogilvie, H. A., and Nakhleh, L. (2019). Unifying gene duplication, loss, and coalescence on phylogenetic networks. bioRxiv https://www.biorxiv.org/content/10.1101/589655v1.

Elworth, R. A. L., Ogilvie, H. A., Zhu, J., and Nakhleh, L. (2019). Advances in computational methods for phylogenetic networks in the presence of hybridization. *Computational Biology*, pages 317–360.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410.

Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.

Gernhard, T. (2008). The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4):769–778.

Gillespie, J. H. (2004). *Population Genetics - A Concise Guide*. JHU Press, second edition edition.

Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., Sarah, G., Santoni, S., David, J., and Ranwez, V. (2019). Pervasive hybridizations in the history of wheat relatives. *Science Advances*, 5(5):eaav9188.

Gogarten, J. P. and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687.

Griffith, F. (1928). The significance of pneumococcal types. *The Journal of Hygiene*, 27(2):113–59.

Hasić, D. and Tannier, E. (2019). Gene tree reconciliation including transfers with replacement is np-hard and fpt. *Journal of Combinatorial Optimization*, 38(2):502–544.

Husnik, F. and McCutcheon, J. P. (2017). Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology*, 16(2):67–79.

Huynen, M. A. and van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Molecular Biology and Evolution*, 15(5):583–589.

Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108.

Jarvis, E. D., Mirarab, S., [...], Gilbert, M. T. P., and Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.

Karev, G. P., Wolf, Y. I., Rzhetsky, A. Y., Berezovskaya, F. S., and Koonin, E. V. (2002). Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evolutionary Biology*, 2(1):18.

Kendall, D. G. (1949). Stochastic processes and population growth. *Journal of the Royal Statistical Society. Series B (Methological)*, 11(2):230–282.

Knowles, D. G. and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Research*, 19(10):1752–1759.

Knowles, L. L., Huang, H., Sukumaran, J., and Smith, S. A. (2018). A matter of phylogenetic scale: Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. *American Journal of Botany*, 105(3):376–384.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biology*, 3(2):research0008.1.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Lerat, E., Daubin, V., Ochman, H., and Moran, N. A. (2005). Evolutionary origins of genomic repertoires in bacteria. *PLoS Biology*, 3(5):e130.

Li, L., Stoeckert Jr., C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189.

Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, 4(11):865–875.

Lynch, M. (2007). *The origins of genome architecture*. Sinauer Associates Sunderland, MA.

Marcet-Houben, M. and Gabaldón, T. (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology*, 13(8):e1002220.

Mendes, F. K., Livera, A. P., and Hahn, M. W. (2019). The perils of intralocus recombination for inferences of molecular convergence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1777):20180244.

Meng, C. and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology*, 75(1):35–45.

Meyer, A. and Schartl, M. (1999). Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Current opinion in cell biology*, 11(6):699–704.

Meyer, A. and Van de Peer, Y. (2005). From 2r to 3r: evidence for a fish-specific genome duplication (fsgd). *Bioessays*, 27(9):937–945.

Miele, V., Penel, S., and Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with silix. *BMC Bioinformatics*, 12(1).

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304.

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer Science & Business Media.

Page, R. D. (2000). Extracting species trees from complex gene trees: Reconciled trees and vertebrate phylogeny. *Molecular Phylogenetics and Evolution*, 14(1):89–106.

Paradis, E. (2016). The distribution of branch lengths in phylogenetic trees. *Molecular Phylogenetics and Evolution*, 94:136–145.

Pett, W., Adamski, M., Adamska, M., Francis, W. R., Eitel, M., Pisani, D., and Wörheide, G. (2019). The role of homology and orthology in the phylogenomic analysis of metazoan gene content. *Molecular Biology and Evolution*, 36(4):643–649.

Philippe, H. and Forterre, P. (1999). The rooting of the universal tree of life is not reliable. *Journal of Molecular Evolution*, 49(4):509–523.

Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D., Bustos-Korts, D., Goué, N., Balfourier, F., et al. (2019). Tracing the ancestry of modern bread wheats. *Nature Genetics*, 51(5):905.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Rannala, B. and Yang, Z. (2013). Improved reversible jump algorithms for bayesian species delimitation. *Genetics*, 194(1):245–253.

Rannala, B. and Yang, Z. (2020). Species delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.5, pages 5.5:1–5.5:18. No commercial publisher | Authors open access book.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765.

Reams, A. B. and Roth, J. R. (2015). Mechanisms of gene duplication and amplification. *Cold Spring Harbor Perspectives in Biology*, 7(2):a016592.

Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., Han, K.-L., Harshman, J., Huddleston, C. J., Kingston, S., et al. (2017). Why do phylogenomic data sets yield conflicting trees? data type influences the avian tree of life more than taxon sampling. *Systematic Biology*, 66(5):857–879.

Reed, W. J. and Hughes, B. D. (2004). A model explaining the size distribution of gene and protein families. *Mathematical Biosciences*, 189(1):97–102.

Richards, E. J., Brown, J. M., Barley, A. J., Chong, R. A., and Thomson, R. C. (2018). Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological? *Systematic Biology*, 67(5):847–860.

Robinson-Rechavi, M. (2020). Molecular evolution and gene function. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.2, pages 4.2:1–4.2:20. No commercial publisher | Authors open access book.

Robinson-Rechavi, M., Marchand, O., Escriva, H., and Laudet, V. (2001). An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. *Current Biology*, 11(12):R458–R459.

Scally, A., Dutheil, J. Y., [...], Tyler-Smith, C., and Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–175.

Scornavacca, C. and Galtier, N. (2017). Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66(1):112–120.

Sidow, A. (1996). Gen(om)e duplications in the evolution of early vertebrates. *Current Opinion in Genetics & Development*, 6(6):715–722.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482.

Springer, M. and Gatesy, J. (2018). Delimiting coalescence genes (c-genes) in phylogenomic data sets. *Genes*, 9(3):123.

Stadler, T. (2009). On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58–66.

Stadler, T. and Steel, M. (2012). Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *Journal of Theoretical Biology*, 297:33–40.

Stadler, T. and Steel, M. (2019). Swapping birth and death: Symmetries and transformations in phylodynamic models. *Systematic Biology*, 68(5):852–858.

Stewart, F. J. (2013). Where the genes flow. *Nature Geoscience*, 6(9):688–690.

Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, E., and Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences of the United States of America*, 109(43):17513–17518.

Szöllősi, G. J. and Daubin, V. (2012). Modeling gene family evolution and reconciling phylogenetic discord. In Anisimova, M., editor, *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*, volume 856 of *Methods in Molecular Biology*, pages 29–51. Springer.

Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V., and Boussau, B. (2015). Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370(1678).

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. (2013a). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6):901–912.

Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013b). Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397.

Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702.

Thompson, E. A. (1975). *Human Evolutionary Trees*. Springer.

Tiley, G. P., Barker, M. S., and Burleigh, J. G. (2018). Assessing the performance of ks plots for detecting ancient whole genome duplications. *Genome Biology and Evolution*.

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, 18(7):411–424.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579.

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, 107(20):9264–9269.

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 213(402-410):21–87.

Zhaxybayeva, O. and Gogarten, P. J. (2004). Cladogenesis, coalescence and the evolution of the three domains of life. *Trends in Genetics*, 20(4):182–187.

Zwaenepoel, A. and Van de Peer, Y. (2019). Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Molecular Biology and Evolution*, 36(7):1384–1404.

Chapter 3.2 Reconciling Gene Trees with Species Trees

Bastien Boussau
Laboratoire de Biométrie et Biologie Évolutive (LBBE)
Université de Lyon, Université Lyon 1, CNRS, Villeurbanne, France
bastien.boussau@univ-lyon1.fr
https://orcid.org/0000-0003-0776-4460

Celine Scornavacca
Institut des Sciences de l'Évolution Université de Montpellier, CNRS, IRD, EPHE
Place Eugène Bataillon 34095
Montpellier Cedex 05, France
celine.scornavacca@umontpellier.fr

Abstract

In the last decade, we witnessed the ascent of *reconciliations* as an important tool to model and study the evolution of gene families. Reconciliations model discordance between gene trees and species trees caused by gene-level processes: duplications, losses and transfers of genes, Incomplete Lineage Sorting among others can be combined to generate a panoply of different models. In this review article, we give an overview of this vast topic by skimming over the different models and methods that have been proposed, and presenting some of their applications in phylogenomics. We also present the pros and cons of these methods and give some directions for future research that we are convinced will enhance their efficiency and use.

How to cite: Bastien Boussau and Celine Scornavacca (2020). Reconciling Gene trees with Species Trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 3.2, pp. 3.2:1–3.2:23. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Gene trees differ from species trees

When studying genome evolution in a set of species, it is often necessary to study the evolution of individual genes that are found across all or most of the species of interest. Assuming genome sequences are available and have been annotated, the first step of such an analysis is to define gene families. Those gene families group together homologous sequences, which are likely to have evolved from a common ancestral sequence. In some cases, there will be exactly one gene per species, in others, some species will be missing the gene, or some species will have more than one copy of the gene. Families with one gene copy per species are typically combined to reconstruct species trees. They can also be subjected to individual phylogenetic analyses, whose steps typically involve aligning the sequences and reconstructing their phylogeny, called a *gene tree*.

When such an analysis is performed, one often observes that many of the reconstructed gene trees do not agree with the (supposedly known) species tree. Here it is important to agree on what is meant by "agreement" between a gene alignment and a reference tree. The measure of disagreement should not be simply topological: a gene tree can differ from the species tree, but not *significantly* so. To conclude that a gene alignment really rejects a particular tree, a statistical test needs to be performed. Significant gene tree/species tree

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 3.2; pp. 3.2:1–3.2:23

A book completely handled by researchers.
No publisher has been paid.

discordances can have two main causes: either they reflect inferential errors or model misspecification, or are due to evolutionary events that have led to truly different topologies between individual gene trees and the species tree.

1.1 Conflicts caused by errors and model inadequacy

Problems can arise at each step of a typical phylogenetic pipeline (see Chapter 2.1 [Simion et al. 2020]). First of all, the sequences themselves can be erroneous: for instance contamination is an issue whose importance has often been underestimated (Simion et al., 2018), and assembly errors are common too. Second, errors can occur during the construction of gene families. It is indeed easy for clustering methods for homology detection to miss a gene if its sequence has diverged a lot compared to the threshold that the user has chosen. Similarly, a gene that contains several protein domains[1] may be incorrectly assigned to a gene family with which it shares one of its domains, but not the others. In both cases, gene trees reconstructed from gene families where such clustering errors have occurred will likely be different from the species tree. One should work towards avoiding such mistakes, for instance by incorporating models of domain fusion/fission during the clustering and alignment steps. Once gene families have been defined, users typically want to extract families of orthologous genes (see Chapter 2.4 [Fernández et al. 2020]), i.e. remove paralogous (=generated by duplication) and xenologous (=generated by transfer) gene copies. However, orthology relationships can be incorrectly inferred, in which case the analysis will be conducted on a group of sequences containing paralogs or xenologs; this may lead to cases as the ones depicted in Figures 2 and 3. Even when orthology is correctly inferred, errors can creep in at the next step, when the sequences are aligned, which may lead to phylogenetic reconstruction errors (see Chapter 2.2 [Ranwez and Chantret 2020]). Finally, our models of sequence evolution are simplistic and for instance very rarely account for dependencies between sites, and heterogeneities of the process across lineages or across sites. Such limitations can introduce errors during phylogenetic reconstruction. For all these reasons, we may observe a high level of discrepancy even in gene families where lateral gene transfers/duplications/losses and reticulate evolution (see Section 1.2) are rare or inexistant. For instance, in birds the amount of discord between gene trees was massive (Jarvis et al., 2015), and similarly in mammals (Scornavacca and Galtier, 2017). In such cases, it is often argued (e.g. in Song et al. 2012; Chapter 3.3 [Rannala et al. 2020]), that a large portion of this incongruence is due to incomplete lineage sorting (ILS, see Figure 4 and the associated section). However, in mammals, using simulations and back-of-the-envelope computations, Scornavacca and Galtier (2017) showed that ILS can only explain a small portion of the incongruence present in the data, see Figure 1. Additionally, in the bird phylogenomic data set, the amount of conflict between trees is larger when the trees are built from exon sequences than when the trees are built from intron sequences, even when the alignment size is taken into account. This result suggests that much incongruence is due to lack of information in the sequences, because exons are typically shorter and more constrained than introns.

1.2 Conflicts caused by biological processes

In this section, we briefly review the biological processes that can generate a gene tree different from the species tree (for more detailed reviews, see Maddison, 1997; Szöllősi et al.,

[1] Protein domains are conserved parts of a given protein sequence that have the characteristics to be able to evolve and exist independently of the rest of the protein chain.

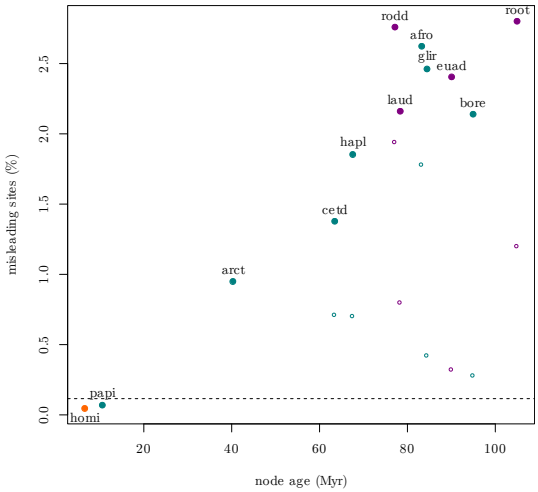


Figure 1 High levels of incongruence present in the OrthoMaM database. Dots correspond to the proportion of parsimoniously-misleading sites for various ancestral nodes in the mammalian phylogeny, and the horizontal line shows the maximal expected percentage of ILS-induced incongruence. Reproduction of Figure 4 in Scornavacca and Galtier (2017), see corresponding paper for more details.

2015). To ease the reading, in our examples we will focus on evolutionary scenarios that generate gene families with exactly one gene per species. However, several of these processes can change copy numbers in a genome, and we will point this out in the description below when relevant.

1.2.1 Gene duplication and gene loss

Gene duplication creates a new copy of a gene, at a different locus in the genome. When gene duplication is followed by gene losses, this may result in a gene tree differing from the species tree even when every extant species ends up with exactly one gene copy. For example, Figure 2(a) depicts a species tree (bold lines) inside of which the evolutionary history of a gene (grey lines) is drawn: first, a speciation happens in R, then the gene is duplicated in the branch leading to U followed by speciations in U and V. This scenario gives rise to seven different genes[2] $a_1, a_2, b_1, b_2, c_1, c_2, d$. Now, if a_2, b_1, c_2 are lost, a_1, b_2, c_1, d may be wrongly identified as orthologs, leading to the tree in Figure 2(b), whose topology differs from that of the species tree (black lines in Figure 2(a)).

1.2.2 Gene transfer, gene conversion

Differences in the topologies of a species tree and a gene tree can be caused by a combination of horizontal gene transfer and gene loss, as shown in Figure 3(a): the copy of the gene present in the ancestral species labelled by V is lost in the species labelled by a and it is replaced by a copy transferred from the species labelled by d. This gives rise to the topology in Figure 3(b), again conflicting with the topology of the species tree.

[2] In the examples of this chapter, we will use the following notation: gene names are associated to small-case letters, possibly subscripted by a number, and species names to capital letters, where genes belong to the species that is associated to the same letter.

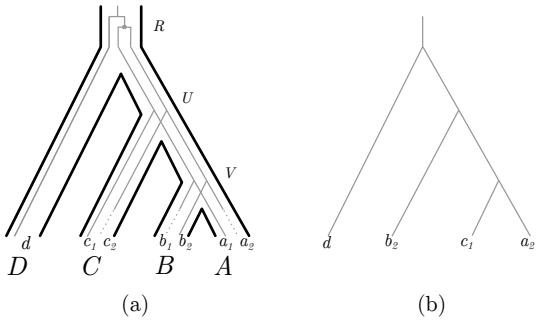


Figure 2 (a) A species tree (bold lines) along with a gene history (grey lines) involving speciations, a gene duplication and gene losses (copies a_2, b_1 and c_2 are lost). (b) The gene tree that may be reconstructed from the data issued from the gene history in (a).

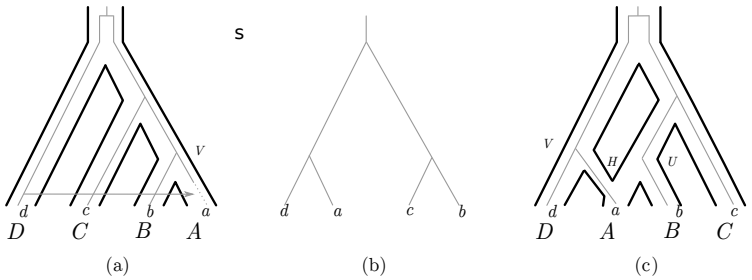


Figure 3 (a) A species tree (bold lines) along with a gene history (grey lines) involving speciations, a gene transfer and a gene loss. (b) The tree that may be reconstructed from a gene alignement resulting from the gene history depicted in (a) or the one depicted in (c). (c) A species network (bold lines) along with a gene history (grey lines) involving a reticulation event.

Most methods assume that a gene transfer adds a new gene copy into the recipient genome. Some documented mechanisms of genetic exchange between genomes, however, involve replacement transfer, i.e., cases when the transferred gene is copied onto an existing homologous copy in the recipient genome by gene conversion. Such an event could result in the topology of Figure 3(b) without the need of any loss event: the recipient genome both gains the transferred copy and loses the resident copy at the same time. In such a case, classical models of gene transfer would typically reconstruct two events when a single one actually occurred (but see Suchard, 2005; Hasic and Tannier, 2017a, for exceptions). Of note, such replacement transfer events, which rely on sequence homology, are only expected to happen between closely related species.

1.2.3 Hybridization

The topology in Figure 3(b) can also be obtained by a scenario involving hybridization, whereby the genome of a descendant lineage is some type of fusion of the genomes of two parental lineages (see Figure 3(c)). Hybridization is increasingly recognized as having an key role in the evolution of some plants and animals, for example in wheat (Glémin et al., 2019) or yeasts (Morales and Dujon, 2012). In this case, the copy of the gene labelled by a is inherited from the ancestral species labelled by V; no copy is inherited from U. Such processes result in dozens to thousands of replacement transfers, occurring through homologous recombination. Because they occur on such different scales and because in some cases hybridization can be associated with a duplication of the genetic material, replacement

transfers and hybridization have classically been modelled differently (Huson et al., 2010; Szöllősi et al., 2015).

1.2.4 Incomplete lineage sorting

Another possible source of conflict between species trees and gene trees is ILS (see Figure 4; Chapters 3.3 and 3.4 [Rannala et al. 2020; Bryant and Hahn 2020]).

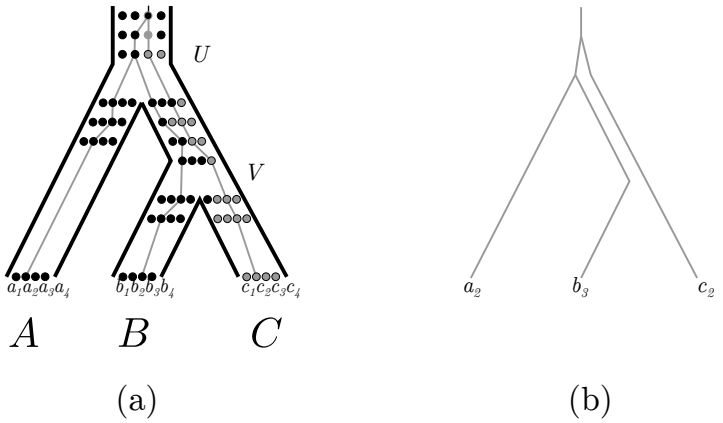


Figure 4 (a) A species tree (bold lines) along with the *coalescent history* of gene variants a_2, b_3 and c_2 (grey lines), descended from alleles present in the common ancestral population U. (b) The gene tree that may be reconstructed from the data issued from the coalescent history depicted in (a).

In this figure, each population contains 4 individuals, each individual is represented with a dot whose color corresponds to a particular allelic form at a given locus, i.e. the black allele and the grey allele. The starting population contains a single allele. First, a mutation leads to a new grey allele at this locus, then a first speciation takes place in U, rapidly followed by a second one in V. As the black and grey alleles still co-exist in V, both alleles still have a chance to be fixed in B and C, so we can obtain the gene tree in Figure 4(b). It is important to realize that such a scenario is entirely neutral: there is no need to assume differences in fitness between the grey and black alleles for such an incongruence to occur. In fact, incomplete lineage sorting can occur whenever speciations occur closely enough that polymorphisms are not entirely "sorted out" between two speciations. Two parameters control the likelihood of such a scenario: first, how fast successive speciations occur. Second, how large the effective population sizes are: in populations with large effective sizes, there will be more polymorphism that stay longer in the population, and therefore more opportunities for ILS.

1.3 Better models to improve gene trees

Discordance between gene trees and species tree, therefore, can result from a variety of processes in combination, and it is often difficult to pinpoint one major cause. One can use better alignment tools and methods to detect contamination and reduce the amount of errors. Other factors can be addressed by improved models of sequence or gene family evolution. For instance, models of sequence evolution that incorporate heterogeneity between sites typically do better than models that treat all sites homogeneously (Lartillot and Philippe, 2004; Wang et al., 2018). Similarly, models that account for the across-lineage variation in

the substitution process should be able to perform better than the most usual models when data show substantial compositional heterogeneity (Boussau and Gouy, 2006; Heaps et al., 2014). Therefore, in practice, efforts should be made to use the most appropriate models of sequence evolution. In this review, we choose to focus on another approach to improve the accuracy of gene and species trees: gene tree-species tree (GTST) models. These models formally describe how a gene family evolves along a species tree and are the focus of the next section.

2 Gene tree-species tree (GTST) models

GTST models describe the evolution of gene trees along species trees, by placing gene duplication (D), transfer (T), loss (L) and conversion (C) events along the gene tree and ILS events along the species tree.

These models can be used in a variety of settings. Historically, people have been using these models in the *reconciliation setting*: the input typically consists of a gene tree and a species tree, and we look for the best scenario that embeds the gene tree inside the species tree, as shown in the figures of the previous section. The aim here may be to estimate the parameters of a given probabilistic GTST model (for instance, the rates of duplication and loss, or population genetic parameters of a model of ILS) or to map events onto the phylogeny (e.g., where the duplications and transfers are placed in the gene history). But these models can also be used to estimate gene trees. In this case, the input is the species tree and data from the genes of interest (e.g. a distribution of gene trees previously reconstructed from the genes, or the gene sequences themselves) and we use an algorithm to look for the gene trees giving the best reconciliation according to some scoring function. The hope is that using a species tree in addition to sequence information will result in an improved estimate of gene trees.

We shall start, after a short digression on parsimony and probabilistic approaches for GTST models (Section 2.1), by reviewing GTST models in the reconciliation setting (Section 2.2) and their extension to account for unsampled species (Section 2.3) and scenario uncertainty (Section 2.4). Finally, we will show how these models can be used to improve the accuracy of gene trees (Section 2.5).

2.1 Parsimony vs probabilistic models

Parsimony approaches have first been used for phylogenetic inference based on morphological or sequence data. Given a set of possible events and a cost for each of them, these methods aim at returning a solution that minimizes a cost function. For phylogenetic inference based only on sequence alignments, the cost function to minimize is usually the sum of the individual costs of events of substitution required to explain the evolution of the sequences along a particular tree topology. For GTST models in the reconciliation setting, the cost function would be the sum of individual costs of events of gene family evolution (typically duplications, losses, transfers or ILS) required to explain the evolution of a gene tree along a particular species tree. In both cases, the costs associated to the events have to be fixed by the user and cannot be estimated.

Probabilistic methods rely on a different type of cost function. Events are associated no longer to costs but to rates, which can be used to compute the probability of various evolutionary scenarios. For example, given rates for all the events considered, probabilistic GTST models in the gene tree-estimation setting (see Section 2.5) enable computing the likelihood of a gene tree, which is proportional to the probability of the gene tree given the

species tree; the posterior probability of a gene tree can be computed by combining the likelihood with prior probability distributions on parameter values. With such probabilistic models, it is possible to estimate the parameters by identifying those maximizing the likelihood of a gene tree. Alternatively, one can integrate over the parameter distribution through Bayesian approaches, and generate the posterior probability of the gene trees and associated parameter values.

2.2 The reconciliation setting

The first and possibly best-known GTST model is the *DL model*:

▶ **Definition 1.** Given a rooted gene tree G and a rooted species tree S whose species contain the genes in G, the evolution of G along S is subject to the following constraints:

1. Speciations are the only possible events shaping species histories;
2. Speciation, duplication (D) and loss (L) are the possible events shaping gene histories;
3. Each speciation in G happens at a speciation in S;
4. L events in G are supposed to happen just after a speciation in S;
5. Each speciation and D event in G gives birth to exactly two genes;
6. The evolution of G along S goes forward in time;
7. Each contemporary gene is a leaf of G and is associated to the corresponding species of S in which this gene is collected.

See the Supplementary Material of Jacox et al. (2016) for a formal and mathematical definition of the model. A *DL reconciliation* is a *plunging* of G in S respecting Def. 1. This plunging can be formalised as a function that maps each node of G onto an ordered sequence of nodes of S.

If we are in the MP framework, we will seek the scenario minimizing the cost $\delta \times |D| + \lambda \times |L|$, where δ and $|D|$, and λ and $|L|$, are respectively the cost and number of events in the scenario for duplications and losses. For this simple model, the best scenario is the Last Common Ancestor (LCA) mapping which can be found in linear time in $|G|$ (Chauve and El-Mabrouk, 2009). In the ML or Bayesian framework, we will compute probabilities of scenarios described via birth-death processes (birth at speciation and duplication events and death at gene losses); in some cases, we may be interested in searching for the scenario with the highest probabilities, in others we can integrate over all scenarios to compute the probability that a given gene tree has evolved along a particular species tree, without explicitly specifying a particular scenario.

This simple model can be made more complex by considering other events shaping the gene history, for example gene transfer. When incorporating gene transfer, Point 2 of Def. 1 becomes:

2. Speciation, duplication (D), loss (L) and transfers (T) between sampled/unsampled species are the possible events shaping gene histories;

We call this model, the *DTL model*. Because transfers necessarily occur between contemporaneous species, transfers to older species are forbidden. However, when some species have not been sampled, it is possible to infer transfers from an ancient donor to a more recent recipient, even if the two species did not live at the same time. This is because a gene may have been transferred to species that are unrepresented in the sample under consideration. If a gene gets transferred to an unrepresented species, stays there for some time, then gets transferred into a lineage ancestral to a sampled species, it will look like a single transfer occurred between two represented lineages that did not live at the same time. This translates as follows:

8. Each T event happens between two coexisting species if transfers are allowed only between sampled species and their ancestors; otherwise, the donor simply has to be older than the recipient.

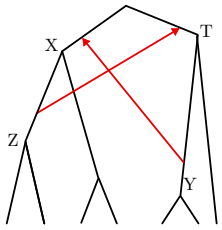


Figure 5 An example of a time-inconsistent scenario: we cannot have node X older than node T and at the same time T older than X.

Now, because of Point 8 of Def. 1, each transfer implies a time constraint between a pair of nodes that may contradict the time constraints implied by other transfers. Computationally, the time constraints implied by gene transfers introduce additional complexity in ensuring Point 6 of Def. 1. Scenarios violating this point are called *time-inconsistent* and can be obtained within a single gene family, especially if it contains several gene copies. An example of a time inconsistent scenario that can be obtained by this latter approach is given in Figure 5. Avoiding time-inconsistent scenarios while preserving optimality is an NP-hard problem (Tofigh et al., 2011), even in the case where we have to reconcile a single binary gene tree with a binary species tree. (Interestingly, one can even make use of time-inconsistent scenarios across gene families to date a species tree, see Section 3.7).

To address this difficulty, two approaches have been used. Either reconciliation is performed against an undated species tree, and one can just hope that time inconsistencies will be rare. Or reconciliation is performed against a dated species tree: all scenarios are therefore consistent with the ages of the nodes of the species tree.

The model can be made even more complex by adding gene conversion (Hasic and Tannier, 2017a), allowing ILS (Vernot et al., 2008; Chan et al., 2017), and accepting unrooted/non binary species and gene trees as input (Górecki and Tiuryn, 2007; Lafond et al., 2016, for instance).

2.3 Transfers to and from the dead

Many models only consider events between branches that have led to extant genes or species. However, much of the past diversity has left no descendant nowadays, and our sampling is necessarily incomplete. For those two reasons, it is likely that a large proportion of the transfers we detect have occurred with a species that has left no descendant among the leaves of our data set. When interpreting reconciled gene histories, it is important to keep this in mind, as this can lead to mistakes. In particular, although transfers necessarily occur between contemporaneous species, the fact that many species have not been sampled means that many donor species will be found on older branches than the sampled recipient species. An example is shown in Figure 6.

Accounting for unsampled species during inference with gene transfer can be done both in a parsimony and a probabilistic framework. In both frameworks, one has to make sure that transfers from a donor on an old branch of the tree can be received by recipients on any branch that is of the same age or more recent than the donor. In the probabilistic framework, one can then model unrepresented species. To this end, an additional modelling

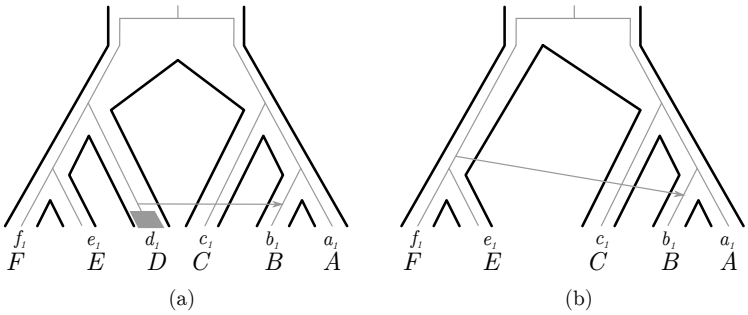


Figure 6 (a) An evolutionary scenario for a gene involving a transfer from the unsampled (or extinct) taxon D. Since D is not present, the transfer is inferred to come from the ancestor of f_1 and e_1, see (b).

layer describing the total number of species living through time must be developed. Szöllősi et al. (2013b) assumed that species evolved according to a Moran model: the number of species vastly outnumbered the extant sampled species but was constant through time. Other models that would allow variations in the number of species through time could be designed. Overall, with this additional layer, GTST models acquire an additional hierarchical level: sequences evolve along gene trees, which evolve with species trees, which are a subsample of an evolving population of species.

In a parsimony model, things are easier. For example, transfers to and from the dead are modelled in ecceTERA as follows: a "dummy" species is added to the species tree, and duplications, losses and transfers to it are free of cost, while transfers from it cost as ordinary transfers. The rational behind this choice is the following: the "dummy" species models all unsampled species thus it can duplicate and loose genes to mimic the number of unsampled species we need; transfers to the "dummy" species are free because they cannot be distinguished from undetected speciation events (undetected because the species became extinct/has not been sampled), while transfers from it are actual transfers.

2.4 Accounting for scenario uncertainty

Whether inference is performed in a parsimony or probabilistic framework, there are cases where several scenarios are nearly equally good descriptors of the evolution of a particular gene tree given a species tree. This would typically occur in a parsimony framework when several scenarios have the same total cost, but this can also occur in a probabilistic framework when several scenarios have very similar likelihoods or probabilities. In such cases, it is important that the reconciliation method returns more than a single scenario, so that the user is fully aware of the uncertainty associated to the inferred events. To put it differently: if a method always returns a single scenario, a user cannot tell if a particular event is necessary to explain a gene tree given a species tree, or if it is just one possible event out of many similarly likely events. If a single scenario is output, a user may well over-interpret an event that is highly uncertain. Another reason for returning several scenarios is depicted in Figure 7: one may be able to choose the preferred scenario among the returned ones using external information, in our example the reconciliation of a neighboring gene family suggesting a single transfer that involved both genes.

This is why almost all reconciliation tools nowadays have opted to show the uncertainty in the scenarios, either providing a measure of support for each event or a posterior distribution of scenarios.

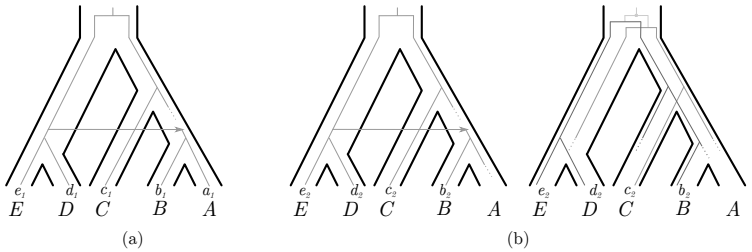


Figure 7 Two sets of reconciliations for two different gene families containing respectively the genes a_1, b_1, c_1, d_1, e_1 and b_2, c_2, d_2, e_2. (a) A scenario for the first family involving a transfer and a loss. (b) Two different scenarios for the second family; the first invokes a transfer and a loss, the second a duplication and four losses. These two scenarios can have the same cost for some vectors of parameters, e.g. if transfers cost four and duplications and losses one, but the first scenario implies a single transfer event that would have moved both genes at the same time.

2.5 Taking into account gene tree uncertainty and improving gene tree accuracy

Beyond uncertainty in the scenario explaining a gene tree given a species tree, there can be huge uncertainty in the gene tree itself. For this reason, it is common practice when inferring gene trees to compute branch support values, for instance through bootstrap (MP, ML), approximations of the bootstrap (ML), or by displaying posterior probabilities on branches (PP). For single gene trees, these support values can be quite low, which shows that there is a lot of uncertainty about the gene tree topology, and which forces interpretations to take branch support into account. Similarly, when inferring gene trees using GTST models, it is very important to take this uncertainty into account.

The two approaches that have been used to take this uncertainty into account in GTST programs are detailed below. In both cases, the program needs a species tree. Then, either the program uses gene sequences to output a distribution of gene trees, or the program takes as input a pre-existing distribution of gene trees, that it will alter and then output. Gene tree estimation based on GTST models can be seen as an effort to come up with an estimation of gene trees that balances between the information provided by the species tree and the information provided by sequence alignments. Hence, there is a choice to be made as to the weights associated to each of these two sources of information: a large weight on the species tree will cause all gene trees to resemble the species tree, while a large weight on sequence information will result in the same trees as obtained using PhyML, RAxML, IQtree, etc, all approaches that only rely on sequence information.

2.5.1 Approaches that take gene sequences as input

Those approaches require using a model of sequence evolution jointly with a GTST model. In a parsimony framework, this requires coming up with a meaningful choice of weights that balances the cost of a substitution with the cost of a duplication, loss, transfer or ILS. In a probabilistic framework, this means that both the parameters of the model of sequence evolution and the parameters of the model of gene family evolution have to be estimated. This creates a challenging problem, because the gene tree also has to be estimated. In addition, computing the likelihood of a gene tree according to an alignment is computationally costly,

which makes these methods time-consuming. An example of this approach is the software jPrIME-DLRS (Sjöstrand et al., 2012). Very recently, two new tools have been proposed for this task: Treerecs (Comte et al., 2019) and GeneRax (Morel et al., 2020), under the DL and DTL model respectively.

2.5.2 Approaches that take gene tree distributions as input

To simplify the inferential problem, several programs rely on input sets of gene trees which have been pre-computed from an alignment of gene sequences. Such sets can be obtained with bootstrap replicates, or thanks to Bayesian inference, in which case the set of trees approximates a probabilistic distribution on gene trees. Typically, such a distribution would be obtained with software for Bayesian gene tree reconstruction, such as MrBayes (Ronquist and Huelsenbeck, 2003), PhyloBayes (Lartillot et al., 2013) and Beast (Suchard et al., 2018). Based on such a set of trees, programs can search for the tree minimizing the cost or maximizing its probability according to a GTST model, or can sample trees according to their probabilities. Relying on a set of trees necessarily comes with a trade-off between accuracy and computational efficiency. If the tree distribution were infinite in size, all the information present in the alignment and exploitable by the model of sequence evolution would be enclosed in the tree distribution, and this approach would be entirely equivalent to the approach described in Section 2.5.1. Of course, the tree distribution has to have a finite size, and therefore cannot describe all the information present in the alignment. The larger the size of the tree distribution, the more accurate the inference will be, but it will also be more costly. In practice, authors have found good trade-offs between accuracy and speed (Edwards et al., 2007; Szöllősi et al., 2013a; Scornavacca et al., 2014), relying notably on the amalgamation idea.

Amalgamation, initially proposed in the parsimony framework by David and Alm (2011) and later formalised and extended to the probabilistic framework (Scornavacca et al., 2014; Szöllősi et al., 2013a), exploits the following ideas. First, a distribution $\mathcal{G}$ of trees induces a distribution of subtrees, which are obtained by cutting the complete trees on each of their internal branches. Second, given a distribution of subtrees, one can mix and match (amalgamate) subtrees with non-overlapping sets of tips to re-create complete trees. Third, provided one computes a few count statistics on the subtrees, it is even possible during this amalgamation step to recapitulate with high accuracy the frequency with which a particular complete tree has been observed in the tree distribution $\mathcal{G}$, i.e. its *conditional clade probability* (CCP). See Figure 8 for an example of amalgamation. Further, this amalgamation trick can also be used to compute the probability of a complete tree that is not present in the input distribution, but can be obtained by amalgamating subtrees found in distinct input trees (Höhna and Drummond, 2011; Larget, 2013). Amalgamation is used in some GTST implementations to integrate over the topological uncertainty associated with the limited information contained in an alignment. Through a single pass of a dynamic programming algorithm, one can integrate over all this uncertainty, either in a parsimony (Scornavacca et al., 2014) or in a probabilistic framework (Szöllősi et al., 2013a).

A note on "overfitting" or shrinkage risk

As pointed out earlier, when using reconciliation to estimate gene trees under a parsimony framework, a reconciled gene tree is a barycentric estimate between a gene tree based on the sequences alone, and a gene tree based on the minimization of the number of e.g. D, T, L events. It can be difficult to place the barycenter. When the reconstructed gene trees tend

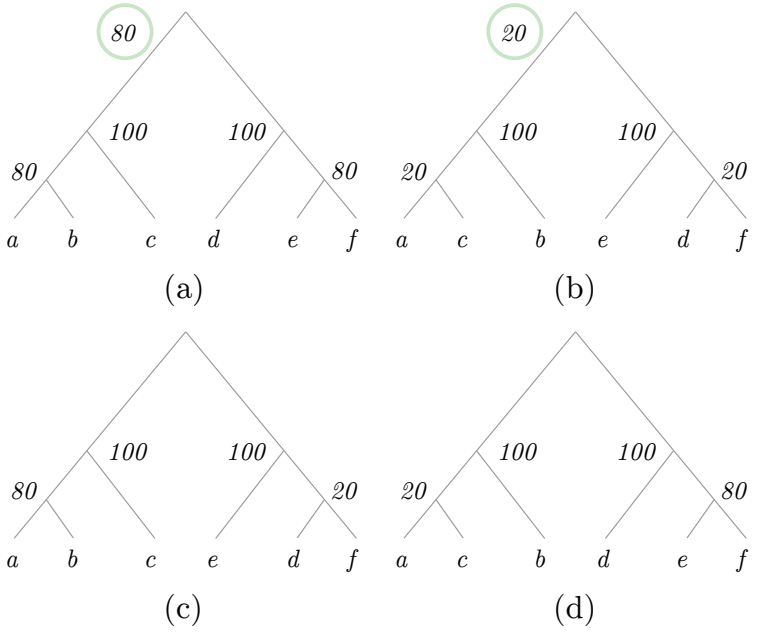


Figure 8 An example of application of the amalgamation principle in the parsimony framework and under the DL model. In (a,b) are presented the trees of our initial distribution, the first present 80 times and the second 20. In (c,d) are depicted the two trees that are not in the initial distribution but can be obtained via amalgamation of the trees in (a,b). The numbers next to the internal nodes show the occurrence of the corresponding clades in the initial distribution. Suppose that our species tree is the tree in (c) and δ, λ and w are respectively the cost of a duplication, the cost of a loss and the weight of the contribution of the sequence alignment to the cost. If the scoring function used is the one presented by Scornavacca et al. (2014), then the costs of the four trees are respectively (a) $\delta + 3\lambda + w \times (2log(80/100) + 2log(100/100))$, (b) $\delta + 3\lambda + w \times (2log(20/100) + 2log(100/100))$, (c) $w \times (log(80/100) + log(20/100) + 2log(100/100))$ and (d) $2\delta + 6\lambda + w \times (log(20/100) + log(80/100) + 2log(100/100))$. (The terms in grey quantify the cost of deviating from the phylogeny preferred by the sequence alignment alone, while the ones in black are the reconciliation cost, which quantify the deviation of the gene tree from the species tree). Under some sets of weights, trees in (c) and (d) will have good scores even though they have never been observed in the input alignment.

to become too similar to the species tree, some authors have said that we are *"overfitting" the species tree*[3]. One could wonder whether the methods described in Section 2.5 suffer from overfitting since they deviate from the signal contained in the sequence alignment to embrace that of the species tree. In practice, Figure 2(a) of Scornavacca et al. (2014) shows that, for most of the existing methods, this is not the case. Also, if needed, users can choose a relative weight of the sequence component with respect to the reconciliation component of the joint score to avoid excessive shrinkage.

For the list of reconciliation tools in the parsimony framework, see Table 1 by Jacox

[3] Actually, using the statistical jargon, it is probably more correct to say that the method "shrinks" the estimate of the gene trees too much, because then each gene tree is "shrunk" to look like the species tree. After all, in regression, overfitting is used to describe a different phenomenon: when a line is drawn that passes through every single point instead of defining a central tendency, which is very similar to a situation where every single gene tree is allowed to have its own tree with its own idiosyncratic topology and parameters.

et al. (2016). This table is only slightly outdated. RANGER-DTL v.2.0 now permits taking as input unrooted and non-binary trees, Mowgli now accounts for ecological traits and ecceTERA for non-binary trees and for ILS, and EUCALYPT provides support values. For a list of published reconciliation tools in the probabilistic framework, see Figure 5 of Szöllősi et al. (2015).

3 Current/future directions

In this section, we will list several lines of research that may help spread the usage of reconciliation tools.

3.1 Mixing parsimony and probabilistic approaches

The major advantage of probabilistic models is that they allow estimating parameters in a proper statistical framework. Maximum likelihood is most appropriate when data is abundant, in which case the maximum of the function is well defined and the parameter values have small confidence intervals. When data is scarce, Bayesian methods –relying on integration– are more robust, because they attempt to integrate over all the uncertainty surrounding parameter values. The two approaches differ in their speed: Maximum Likelihood is usually faster than Bayesian integration. But both are typically much slower than parsimony approaches, which have no parameter to optimize (see Table 1).

	ecceTERA	ALE_ML	ALE_MCMC
MIN	0.01	0.03	0.03
Q_1	0.19	12.40	90.75
MEDIAN	0.25	16.41	126.96
Q_3	0.32	23.58	184.80
MAX	1.10	520.41	941.39

Table 1 Running times in seconds for three different methods –ecceTERA (Jacox et al., 2016), ALE_ML (Szöllősi et al., 2013a) and ALE_MCMC (Szöllősi and Boussau, 2018), respectively a parsimony, ML and Bayesian method– on a dataset of 1099 homologous gene families present in 36 cyanobacterial genomes (the mean number of genes per family in the dataset is 36.66, the largest family has 114 genes and the smallest 21 genes, see Szöllősi et al., 2013a, for more details).

Although parsimony and probabilistic approaches differ in their treatment of the parameters, they share important similarities in the algorithms used to compute the cost functions. In both cases, the algorithms often involve dynamic programming performed during tree traversals. In the field of gene tree reconstruction based on gene sequences only, many efficient algorithms combine probabilistic and parsimony approaches to benefit from their distinct qualities. Probabilistic models are used to operate in a sound statistical framework, which offers guarantees about the inference. Parsimony is used to speed up the algorithms and provide pretty good solutions fast. For instance, RAxML relies on parsimony to generate starting trees (Stamatakis, 2006), and MrBayes 3.2 has a "parsimony-biased" SPR move (Ronquist et al., 2012). So far, reconciliation methods have not completely merged the parsimony and likelihood frameworks, but we believe there would be much opportunity for doing so. As described above for approaches that rely on sequence information only, new algorithms could be developed that would benefit from the speed of the parsimony cost function to more efficiently optimize or integrate a probabilistic cost function, *i.e.* the

likelihood or posterior probability of a gene tree. Alternatively, one could use algorithms such as importance sampling, where one samples parameter values from a quick-to-compute distribution, computes the probability of those parameter values using a more complex distribution, and finally applies some simple reweighing of the initial distribution to obtain an unbiased sample from the more complex distribution. More specifically for GTST models, one could sample gene trees from a GTST parsimony model, compute their probability according to a sophisticated probabilistic model while sampling values of the parameters of the probabilistic model that were not present in the parsimony model, and finally reweigh the initial sample to obtain a sample of gene trees according to the sophisticated probabilistic model. For such a solution to be useful, sampling from the parsimony model should produce gene trees that have good probabilities, and the sampling of the parameters of the probabilistic model should not be computationally too costly.

3.2 Using reconciliation to estimate species trees

GTST models have also been used to score species trees. When combined with an algorithm that explores species tree topologies, this allows sampling species trees or finding the best species tree according to a particular GTST model based on a large number of gene families from the genomes of interest. Such an approach therefore potentially relies on a huge amount of data, and can bypass the need to identify families of orthologous genes, since D, T, L models can handle families with more than one gene copy per species (*i.e.* homologous genes). When the focus is on the species tree, it can be unnecessary to obtain individual gene family reconciliation scenarios. In fact, it becomes desirable to integrate over all possible reconciliation scenarios between the gene family and the species tree, and obtain a probability or a score that such a gene family would have evolved along a given species tree, irrespective of what particular events were involved, and on which branches they occurred. Such a probability or score is easy to compute using dynamic programming algorithms, which can integrate over or maximize a score or probability without any change to their structure. This approach has for instance been used in PHYLDOG (Boussau et al., 2013). Alternatively, in MCMC algorithms that sample parameters from probabilistic models, integration over all reconciliation scenarios can also be performed by sampling different scenarios at each step of the MCMC. The choice between these two methodologies would come down to which is most efficient to run, since they are equally difficult to implement. To speed up species tree inference in an ML framework, Ullah et al. (2015) have proposed a two-step algorithm whereby a set of candidate species trees built from families of orthologous genes is evaluated according to a DL probabilistic model. An approach combining fast parsimony methods with probabilistic inference as discussed in the previous section could also be a valuable option here.

3.3 Gene conversion vs gene transfer

As pointed out in the first section, gene conversion is not modelled well by most models of gene transfer, because it is a type of replacement transfer. To the exception of the model by Suchard (2005), all the models of transfer consider that a gene transfer adds a copy of a gene to a recipient genome. In those models, gene conversion has to be modelled by two events, a gene transfer and a gene loss. Gene families in which events of gene conversion have occurred will therefore have very unlikely scenarios according to most models of gene transfer, as they will require twice the number of events that actually occurred. In such cases, the barycentric estimation of the gene tree could be off: the GTST model will push

too much towards gene trees that resemble the species tree, because any difference costs twice as much as it should. For this reason, modelling accurately gene families with events of gene conversion would require developing a new model. Such a development is difficult: replacement transfers or gene conversions introduce dependencies between otherwise independent branches of the species tree, which breaks dynamic programming algorithms, the workhorses of all GTST methods. The model of replacement transfer (Suchard, 2005) uses a different type of algorithm, but is extremely limited in the size of the data sets it can handle. Short of developing a better model of gene conversion, users of GTST methods have two options. First, they could try to mitigate the impact of such a model misspecification by tweaking the parameters controlling the penalty or the probability of transfers and losses. By making transfers and losses cost less, or be more probable, scenarios involving gene conversions will be less unlikely. Second, they could use network methods, which are typically designed to describe cases of hybridization or genome-wide reticulation (see section 3.6). This solution requires setting parameters of reticulation, which are usually estimated using genome-wide data and will not be well estimated using a single gene family.

Very recent advances on modelling gene conversion as a single event in the parsimony framework (Hasic and Tannier, 2017a,b) give us hope that gene conversion will be soon better modelled by reconciliation methods.

3.4 Reconciliation of all processes together

It would be very convenient to have a method that can handle all the processes that make gene trees differ from species trees. Such a method could identify which processes are at work in a particular gene family, on particular branches of the species tree. Duplications, transfers and losses have already been merged together, both in parsimony and in probabilistic models (Szöllősi et al., 2013a; Scornavacca et al., 2014; Sjöstrand et al., 2014, among others); the same holds for combining DL with incomplete lineage sorting (Rasmussen and Kellis, 2012; Wu et al., 2014, even though with some simplistic assumptions on how gene duplication and ILS interact). These latter models add an extra hierarchical layer on top of typical DL models: sequences evolve along a gene tree, which evolves according to a coalescent process along a locus tree, which evolves according to a birth-death process along the species tree. This model could be extended to account for gene transfers as well. In the parsimony framework, similar efforts have been attempted, the most complete model being the one of (Chan et al., 2017), combining DTL with incomplete lineage sorting (this method does also oversimplified assumptions on the interaction between duplication and ILS).

As evoked in Section 3.3, so far no model has combined replacement transfers with typical models of DTL, because the classical dynamic programming algorithms cannot be used when replacement transfers are included.

3.5 Reconciliation of several loci together

Gene trees can be reconstructed one by one using GTST models. When probabilistic models are used, this typically involves estimating parameters of the GTST models, such as the rates of various events. This can be difficult on single gene families, where the amount of information is necessarily limited. In such cases, rates can be mis-estimated, and reconciliation scenarios can be wrong. To improve rate estimation, information could be gathered across gene families, by performing joint estimation of the rates. This would reduce stochastic errors, but could result in over-shrinkage effects, where outlier loci with atypical parameter values would be constrained to share the parameter values estimated on other gene families.

A balance between gene-based and genome-based estimation of the rates could be obtained by borrowing ideas from models of rate heterogeneity across sites (Yang, 1994). Such models estimate an average rate of evolution across all sites, but allow variation around the average by estimating an additional variance parameter. There could be additional variance parameters to account for variation in the rates of D, T, L or incomplete lineage sorting across gene families.
Moreover, some events of gene family evolution affect more than one gene family at a time. For instance, duplication, transfer and loss events can affect a segment of the genome that contains several genes. To identify such events it is desirable to analyse several gene families jointly and reconstruct joint scenarios. This, however, is difficult to do. First, events affecting segments of the genome may involve different numbers of genes in different lineages of the species tree. Therefore either an arbitrary choice is made on the number of jointly analyzed genes, or one has to come up with a method that will find the appropriate number for each branch, or assume that co-evolving genes co-evolve throughout their history. A method that identifies co-evolving genes per branch and that uses this information to reconstruct their history would be difficult to design as it would need to consider a vast number of possibilities. Chan et al. (2013) adopted a two-step approach: first, gene families were reconciled individually, and then probabilities that two gene families co-evolved over their entire history were computed based on the individual evolutionary scenarios. This approach was able to recover co-evolving genes in a simulation.

Second, genomes do not remain collinear throughout their evolution: two genes could be neighbors in one genome, but far from each other in another. Therefore either one focuses on the few genes whose relative positions have remained constant throughout their evolution, or one has to use a model of how genes move across the genome. The latter approach has been used in a series of papers that describe models using synteny information to reconstruct ancestral genome structures (Bérard et al., 2012; Patterson et al., 2013; Semeria et al., 2015) that are described in more depths in Chapter 2.5 (Tannier et al. 2020). These methods take as input a (dated) species tree[4], a set of reconciled gene trees and the set of extant adjacencies. The output is a set of ancestral adjacencies that, combined, give the ancestral genome structures. Their underlying model permits the creation and the loss of adjacencies and looks for an adjacency history that is compatible with the given reconciled gene trees and minimizes the number of adjacency creations and adjacency losses. Obviously, it would be interesting to reconcile the gene trees and, at the same time, minimize adjacency creations and losses, but the problem seems to be very complex and it has been conjectured to be NP-hard. To this day, the complexity of this problem is still open, even if the proof for a related problem presented in a very recent paper (Delabre et al., 2018) could shed some light.

Third, the space of all possible events then increases: for instance, in addition to D, T, L of single gene events, one needs to include events of D, T, L of at least pairs of genes. Very recently, methods to take into account duplications and losses of several genes at once (the so-called *segmental* duplications and losses) have been proposed in the parsimony framework (Delabre et al., 2018; Dondi et al., 2018). Overall, the development of models that consider the coevolution of several genes at a time is quite complicated and results in algorithms of high complexity.

[4] The species tree needs to be dated only in Patterson et al. (2013).

3.6 Reconciliation with a species network

When the species phylogeny includes reticulation events such as hybridizations, we talk about *species networks*. Species network inference will not be reviewed in this book but we refer to the excellent recent reviews of Degnan (2018) and Elworth et al. (2019). Here we only aim at highlighting the similarities between some of the approaches to infer networks and the reconciliation methods presented in this chapter. For example, the MDC (Minimizing Deep Coalescence, e.g. Yu et al., 2011), is implicitly based on a reconciliation model in a parsimony framework permitting speciation and ILS at the gene level, and speciation and hybridization at the species level. In more details, gene trees evolve *inside* the network via speciations (hybrid or not)–giving a set of possible trees *associated* to the network– and a given gene tree can be reconciled with each of these trees via speciation and ILS. The best reconciliation w.r.t. the network is then the most parsimonious one over all possible scenarios and trees inside the network. The same rationale underlies similar methods in the probabilistic framework (e.g. Yu et al., 2014; Zhang et al., 2017). These latter models are very time consuming. Other models explicitly extended the reconciliation model to species networks, e.g. To and Scornavacca (2015) for the DL model and Scornavacca et al. (2017) for the DTL model. These methods do not take ILS into account yet. Roughly speaking, they consist in replacing Point 1 of the model described in Section 2.2 with:

1. Speciations and hybridizations are the only possible events shaping species histories;

These explicit models are extremely scalable but also very recent and they have yet to prove their worth.

3.7 Improved dating with reconciliation

Dating a species phylogeny is a difficult endeavour that usually involves using fossil calibrations with relaxed clock models of the rate of sequence evolution. Although much work has been devoted to such inferences, dating a tree remains difficult because disentangling rate and time is fundamentally very hard. Basically, one needs to estimate a rate and a length of time per branch, all this based on an estimate of their product, the branch length. One possibility to improve the inference of dated phylogenies would be to include other events than just events of substitution. In particular, reconciliations notably allow identifying events of D, T, L and placing them on branches of the species phylogeny. One could use these estimated numbers of events on each branch of the phylogeny to better disentangle branch length and rates of evolution, because time will affect in a similar way events of substitution and D, T, L events, while the rates of those events may be partially uncorrelated. By using more events, one could thus better estimate dated phylogenies.

Another approach to improve dating based on reconciliations is to use individual transfer events on their own. Transfer events necessarily occur between contemporaneous species. Given that species sampling is necessarily incomplete, transfers can only indicate that the ancestor of a donor species is necessarily older than the descendant of a recipient species (see Figure 5). Yet, such an information tells us something on the relative age of two nodes in the species tree, and the detection of a transfer event can be translated into a time constraint between nodes of a phylogeny if the donor and a recipient can be identified. When a set of transfers is detected by interpreting the phylogenetic discordance between a gene tree and a species tree, the set of all deduced time constraints can be used to rank the species tree, i.e. order totally its internal nodes. Several genes can give contradicting information that needs to be sorted out (Chauve et al., 2017); still, it seems that transfers can be successful in providing insights into the timing of diversification of clades across the

tree of life (Davín et al., 2018). Combining this transfer-based information with relaxed clock models of sequence evolution and with fossil calibration could result in much more accurate dates for the tree of life, even in taxa/epochs where the fossil record is scant.

3.8 RecPhyloXML and reconciliation visualisation

Recently, a common effort of a consortium of researchers involved in reconciliation-related software development resulted in the introduction of an integrative and flexible format to describe reconciliations (Duchemin et al., 2018). The format is based on grammars extending the PhyloXML format, which is aimed at representing annotated trees in XML. Roughly, in RecPhyloXML the species tree and the reconciled gene tree are described in PhyloXML. Then, each node of the reconciled gene tree is associated to a set of nodes of the species tree via tags that specify the type of event (duplication, transfer, etc), the support and geographical annotations, for instance.

We are confident that RecPhyloXML will ease the development of generic software permitting to visualise and compare reconciliations, such as Sylvix (Chevenet et al., 2015) and the web interface `http://phylariane.univ-lyon1.fr/recphyloxml/recphylovisu`. In turn, this will help the spread of the usage of the reconciliation tools tremendously. The practitioners will not need anymore to spend hours trying to understand the different outputs of different reconciliation tools; they will be able to easily compare them and choose the best software for their data.

3.9 Impact of reconciliation on other steps of the phylogenomic pipeline

We have seen that GTST models can improve gene tree inference. It is known that using better guide trees, e.g. phylogenetic trees guiding how pairwise alignments are combined to obtain a multiple alignment, improves alignment inference (Liu et al., 2009). Therefore it seems likely that using a species tree with a GTST model could improve alignment inference, though not for all instances (see Chapter 2.3 [Ranwez and Delsuc 2020]). Similarly, clustering gene sequences into homologous gene families could benefit from the information coming from the species tree. Clustering methods typically rely on fixed thresholds to include a sequence into a cluster: if a sequence is similar enough to one or several sequences of a cluster, it is included in the cluster, if not, it is excluded. The reliance on such a fixed threshold could be relaxed with some knowledge of the structure of the species tree and of its branch lengths. For instance, one could normalize scores or adapt the threshold to the phylogenetic distance between the considered species as has been done by Emms and Kelly (2015). One could go even further by using reconciliations with models of gene duplication, transfer and loss. Such models would penalize families that are represented in a patchy way in the species phylogeny, in particular in clades where transfer rates are low. Indeed, patchy families would require large numbers of duplications and losses, which would be associated to a low reconciliation score. If we use the species tree for the clustering step, one could develop an iterative approach where a first clustering is used to generate a species tree, which is then used to re-cluster the genes, taking into account the species tree.

4 Conclusion

Several reconciliation methods have been developed these past few years. These allow inferring better gene trees and quantifying the impact of gene-level processes on genome evol-

ution. Progress remains to be made to integrate more processes together (e.g. DTL with ILS, hybridization and gene conversion for multiple genes) in a single inferential method, to allow for correlations between gene histories, and to speed up methods for species tree inference. But reconciliation methods can already contribute to a better understanding of many processes of molecular evolution, and could improve the accuracy of several steps of our phylogenomic pipelines.

References

Bérard, S., Gallien, C., Boussau, B., Szöllősi, G. J., Daubin, V., and Tannier, E. (2012). Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, 28(18):i382–i388.

Boussau, B. and Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution. *Systematic Biology*, 55(5):756–768.

Boussau, B., Szöllősi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330.

Bryant, D. and Hahn, M. W. (2020). The concatenation question. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.4, pages 3.4:1–3.4:23. No commercial publisher | Authors open access book.

Chan, Y.-b., Ranwez, V., and Scornavacca, C. (2013). Reconciliation-based detection of co-evolving gene families. *BMC bioinformatics*, 14(1):332.

Chan, Y.-b., Ranwez, V., and Scornavacca, C. (2017). Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. *Journal of theoretical biology*, 432:1–13.

Chauve, C. and El-Mabrouk, N. (2009). New perspectives on gene family evolution: losses in reconciliation and a link with supertrees. In *Annual International Conference on Research in Computational Molecular Biology*, pages 46–58. Springer.

Chauve, C., Rafiey, A., Davin, A., Scornavacca, C., Veber, P., Boussau, B., Szollosi, G., Daubin, V., and Tannier, E. (2017). Maxtic: Fast ranking of a phylogenetic tree by maximum time consistency with lateral gene transfers. *Peer Community in Evolutionary Biology*.

Chevenet, F., Doyon, J.-P., Scornavacca, C., Jacox, E., Jousselin, E., and Berry, V. (2015). Sylvx: a viewer for phylogenetic tree reconciliations. *Bioinformatics*, 32(4):608–610.

Comte, N., Morel, B., Hasic, D., Guéguen, L., Boussau, B., Daubin, V., Penel, S., Scornavacca, C., Gouy, M., Stamatakis, A., Tannier, E., and Parsons, D. P. (2019). Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. *bioRxiv*, https://www.biorxiv.org/content/early/2019/10/11/782946.

David, L. A. and Alm, E. J. (2011). Rapid evolutionary innovation during an archaean genetic expansion. *Nature*, 469(7328):93.

Davín, A. A., Tannier, E., Williams, T. A., Boussau, B., Daubin, V., and Szöllősi, G. J. (2018). Gene transfers can date the tree of life. *Nature ecology & evolution*, 2(5):904.

Degnan, J. H. (2018). Modeling Hybridization Under the Network Multispecies Coalescent. *Systematic Biology*, 67(5):786–799.

Delabre, M., El-Mabrouk, N., Huber, K. T., Lafond, M., Moulton, V., Noutahi, E., and Castellanos, M. S. (2018). Reconstructing the history of syntenies through super-reconciliation. In *RECOMB International conference on Comparative Genomics*, pages 179–195. Springer.

Dondi, R., Lafond, M., and Scornavacca, C. (2018). Reconciling Multiple Genes Trees via Segmental Duplications and Losses. In Parida, L. and Ukkonen, E., editors, *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*, volume 113 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 5:1–5:16, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Duchemin, W., Gence, G., Arigon Chifolleau, A.-M., Arvestad, L., Bansal, M. S., Berry, V., Boussau, B., Chevenet, F., Comte, N., Davín, A. A., et al. (2018). Recphyloxml-a format for reconciled gene trees. *Bioinformatics*, 1:7.

Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14):5936–5941.

Elworth, R. A. L., Ogilvie, H. A., Zhu, J., and Nakhleh, L. (2019). Advances in computational methods for phylogenetic networks in the presence of hybridization. In Warnow, T., editor, *Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret*, pages 317–360. Springer International Publishing, Cham.

Emms, D. M. and Kelly, S. (2015). Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology*, 16(1):157.

Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.

Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., Sarah, G., Santoni, S., David, J., and Ranwez, V. (2019). Pervasive hybridizations in the history of wheat relatives. *Science Advances*, 5(5):eaav9188.

Górecki, P. and Tiuryn, J. (2007). Inferring phylogeny from whole genomes. *Bioinformatics*, 23(2):e116–e122.

Hasic, D. and Tannier, E. (2017a). Gene tree reconciliation including transfers with replacement is hard and fpt. *arXiv preprint arXiv:1709.04459*.

Hasic, D. and Tannier, E. (2017b). Gene tree species tree reconciliation with gene conversion. *arXiv preprint arXiv:1703.08950*.

Heaps, S. E., Nye, T. M., Boys, R. J., Williams, T. A., and Embley, T. M. (2014). Bayesian modelling of compositional heterogeneity in molecular phylogenetics. *Statistical Applications in Genetics and Molecular Biology*, 13(5):589–609.

Höhna, S. and Drummond, A. J. (2011). Guided tree topology proposals for bayesian phylogenetic inference. *Systematic Biology*, 61(1):1–11.

Huson, D. H., Rupp, R., and Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press.

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., and Scornavacca, C. (2016). eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Alfaro-Núñez, A., Narula, N., Liu, L., Burt, D., Ellegren, H., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T. P., Zhang, G., and Consortium, T. A. P. (2015). Phylogenomic analyses data of the avian phylogenomics project. *GigaScience*, 4(1):s13742–014–0038–1.

Lafond, M., Noutahi, E., and El-Mabrouk, N. (2016). Efficient non-binary gene tree resolution with weighted reconciliation cost. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 54. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Larget, B. (2013). The estimation of tree posterior probabilities using conditional clade probability distributions. *Systematic Biology*, 62(4):501–511.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). Phylobayes mpi: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, 62(4):611–615.

Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., and Warnow, T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934):1561–1564.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

Morales, L. and Dujon, B. (2012). Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. *Microbiology and Molecular Biology Reviews*, 76(4):721–739.

Morel, B., Kozlov, A. M., Stamatakis, A., and Szöllősi, G. J. (2020). Generax: A tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss. *bioRxiv, https://www.biorxiv.org/content/early/2020/02/20/779066*.

Patterson, M., Szöllősi, G., Daubin, V., and Tannier, E. (2013). Lateral gene transfer, rearrangement, reconciliation. *BMC bioinformatics*, 14(15):S4.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Ranwez, V. and Chantret, N. (2020). Strengths and limits of multiple sequence alignment and filtering methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.2, pages 2.2:1–2.2:36. No commercial publisher | Authors open access book.

Ranwez, V. and Delsuc, F. (2020). Accurate alignment of (meta)barcoding datasets using macse. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.3, pages 2.3:1–2.3:31. No commercial publisher | Authors open access book.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome research*, pages gr–123901.

Ronquist, F. and Huelsenbeck, J. P. (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.

Scornavacca, C. and Galtier, N. (2017). Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66(1):112–120.

Scornavacca, C., Jacox, E., and Szöllősi, G. J. (2014). Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, 31(6):841–848.

Scornavacca, C., Mayol, J. C. P., and Cardona, G. (2017). Fast algorithm for the reconciliation of gene trees and lgt networks. *Journal of theoretical biology*, 418:129–137.

Semeria, M., Tannier, E., and Guéguen, L. (2015). Probabilistic modeling of the evolution of gene synteny within reconciled phylogenies. *BMC bioinformatics*, 16(14):S5.

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J. C., Manuel, M., Philippe, H., and Telford, M. J. (2018). A software tool ?croco?detects pervasive cross-species contamination in next generation sequencing data. *BMC biology*, 16(1):28.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Sjöstrand, J., Sennblad, B., Arvestad, L., and Lagergren, J. (2012). Dlrs: gene tree evolution in light of a species tree. *Bioinformatics*, 28(22):2994–2995.

Sjöstrand, J., Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B., and Lagergren, J. (2014). A bayesian method for analyzing lateral gene transfer. *Systematic biology*, 63(3):409–420.

Song, S., Liu, L., Edwards, S. V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 109(37):14942–14947.

Stamatakis, A. (2006). Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Suchard, M. A. (2005). Stochastic models for horizontal gene transfer: taking a random walk through tree space. *Genetics*.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evolution*, 4(1):vey016.

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. (2013a). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6):901–912.

Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013b). Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397.

Szöllősi, G. J. and Boussau, B. (2013–2018). ALE. https://github.com/ssolo/ALE.

Szöllősi, G. J., Tannier, E., Daubin, V., and Boussau, B. (2015). The inference of gene trees with species trees. *Systematic Biology*, 64(1):e42–e62.

Tannier, E., Bazin, A., Davín, A. A., Guéguen, L., Bérard, S., and Chauve, C. (2020). Ancestral genome organization as a diagnosis tool for phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.5, pages 2.5:1–2.5:19. No commercial publisher | Authors open access book.

To, T.-H. and Scornavacca, C. (2015). Efficient algorithms for reconciling gene trees and species networks via duplication and loss events. *BMC genomics*, 16(10):S6.

Tofigh, A., Hallett, M., and Lagergren, J. (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(2):517–535.

Ullah, I., Parviainen, P., and Lagergren, J. (2015). Species tree inference using a mixture model. *Molecular Biology and Evolution*, 32(9):2469–2482.

Vernot, B., Stolzer, M., Goldman, A., and Durand, D. (2008). Reconciliation with non-binary species trees. *Journal of Computational Biology*, 15(8):981–1006.

Wang, H.-C., Minh, B. Q., Susko, E., and Roger, A. J. (2018). Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Systematic Biology*, 67(2):216–235.

Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., and Kellis, M. (2014). Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome research*, 24(3):475–486.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution*, 39(3):306–314.

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46):16448–16453.

Yu, Y., Than, C., Degnan, J. H., and Nakhleh, L. (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60(2):138–149.

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2017). Bayesian inference of species networks from multilocus sequence data. *Molecular biology and evolution*, 35(2):504–517.

Chapter 3.3 The Multi-species Coalescent Model and Species Tree Inference

Bruce Rannala
Department of Evolution and Ecology, University of California Davis
One Shields Avenue, Davis CA USA
brannala@ucdavis.edu
https://orcid.org/0000-0002-8355-9955

Scott V. Edwards
Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University
Cambridge, MA 02138, USA
sedwards@fas.harvard.edu
https://orcid.org/0000-0003-2535-6217

Adam Leaché
Department of Biology & Burke Museum of Natural History and Culture, University of Washington
Seattle, WA 98195-1800, USA
leache@uw.edu
https://orcid.org/0000-0001-8929-6300

Ziheng Yang[1]
Department of Genetics, Evolution and Environment, University College London
London WC1E 6BT, United Kingdom
z.yang@ucl.ac.uk
https://orcid.org/0000-0003-3351-7981

Abstract

The multispecies coalescent (MSC) is an extension of the single-population coalescent model of population genetics to the case of multiple species. The MSC naturally accommodates speciation events (with subsequent genetic isolation between species) and the coalescent process within each species. It provides a framework for analysis of multilocus genomic sequence data from multiple species in a number of inference problems including species tree estimation, accounting for ancestral polymorphism and deep coalescence. Within this framework, the genealogical fluctuations across genes or genomic regions (and the gene tree/species tree conflicts that may result) are not seen as a problem but rather as a source of information for estimating important parameters such as species divergence times, ancestral population sizes, and the timings, directions, and intensities of cross-species introgression or hybridisation events. This chapter outlines the basic theory of the MSC and its important applications in analysis of genomic sequence data, describing the most widely-used full-likelihood and heuristic methods of species tree estimation. We discuss several active areas of research in which we predict future developments will occur, including inference of introgression events on a species phylogeny.

How to cite: Bruce Rannala, Scott V. Edwards, Adam Leaché, and Ziheng Yang (2020). The Multi-species Coalescent Model and Species Tree Inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 3.3, pp. 3.3:1–3.3:21. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

[1] Z.Y. is supported by a Biotechnological and Biological Sciences Research Council grant (BB/P006493/1).

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 3.3; pp. 3.3:1–3.3:21
A book completely handled by researchers.
No publisher has been paid.

1 Introduction

The need by scientists and the public for robust phylogenies and a complete Tree of Life grows every day. Phylogenies are a fundamental building block of evolutionary biology. They provide a detailed genealogical "map" that has applications in a variety of fields such as biogeography, molecular evolution, pathogen evolution, and comparative genomics. In recent years, our ability to infer phylogenies has grown dramatically, not only because of technical advances in high-throughput DNA sequencing, but also through theoretical advances (Boussau et al., 2013; Bravo et al., 2019; Liu et al., 2009, 2015; Rannala and Yang, 2017, 2008). Of the many types of theoretical advances that have been made in the last 20 years, this chapter will focus on the application of the multispecies coalescent model (MSC) to phylogenetic inference. We regard this as one of the most important new directions for phylogenetics since DNA sequencing became more widespread among systematists in the late 1980s.

When the polymerase chain reaction (PCR) became widely available in the late 1980s, population geneticists and evolutionary biologists immediately began estimating phylogenetic trees with DNA sequences (Kocher et al., 1989). Molecular systematics of course goes back even further, but molecular cloning of individual genes was laborious. Within-species studies of gene trees first became possible with the advent of restriction enzymes and their application to DNA diversity in the late 1970s (Brown et al., 1982; Wilson et al., 1985). The shift from allozymes and protein polymorphisms to DNA differences had profound impacts on the evolutionary biology community, not only technically but also because of the insights that were provided to empiricists and theoreticians (Avise, 1994). Whereas allozyme electrophoresis could allow one to tell different alleles in a given species apart, DNA differences allowed one to measure the evolutionary or genetic distance between alleles (Avise et al., 1979). This improved precision led population genetics and phylogenetics into a wholly new territory.

Early investigations of gene trees in closely related populations and species quickly revealed that the gene tree of alleles from different populations did not always correspond to the species tree (Avise et al., 1987). One of the most common reason for this discordance is now well understood – the failure of alleles to coalesce as one moves backward in time toward successive speciation events, or, thinking forward in time, the failure of genetic drift to "sort" alleles into their descendant populations fast enough before the next speciation event. Adopting a forward-time definition, this phenomenon was dubbed "incomplete lineage sorting" by Avise and, taking a backward time perspective, "deep coalescence" by Maddison (1997). Gene tree-species tree discordance can be caused by other biological processes such as gene duplication, introgression or horizontal gene transfer (Nichols, 2001; Edwards, 2009; Szollosi et al., 2015), but these are not inherent to population divergences in the same fundamental way that the coalescent process is because the coalescent operates in all finite populations whereas the other processes are not always present.

Avise also formalized the distinction between a gene tree and a species tree. The concept of a species tree is synonymous with phylogeny and had, of course, been fundamental to evolutionary biology since Darwin's On the Origin of Species was published in 1859 (Darwin, 1859). However, it was empirical studies of gene trees in natural populations that drove home the distinction between a gene tree and the species tree that generated it (Hare, 2001). In the early days of DNA sequencing, and frequently even today, researchers refer to the gene tree as the species tree, or use methods, such as concatenation, that assume that the two are the same (see Chapter 3.4 [Bryant and Hahn 2020]). Although the distinction between gene trees and the species tree has been appreciated for decades, computational methods for estimating the species tree accommodating gene tree discordance have only been available since about 2006.

The gene tree-species tree mismatch probability in the case of three species was derived by

Hudson (1983). The mismatch probability was used to estimate the population sizes for the human-chimpanzee common ancestor (Takahata et al., 1995). The probabilities of gene tree topologies (typically assuming one sequence from each species) given a species tree was further studied by Pamilo and Nei (1988) and more recently by Rosenberg (2002), Degnan and Salter (2005), Degnan and Rosenberg (2006) and Wu (2012, 2016), who developed algorithms for automatic calculation of such probabilities. The most well-known result from this line of research is the existence of the so-called anomaly zone, the zone of species tree and parameter values for which the most probable gene tree has a different topology from the species tree. The full probability distribution of gene trees with branch lengths (coalescent times) for an arbitrary species tree – the multispecies coalescent model - was first fully described by Rannala and Yang (2003). This forms the basis for exact or full-likelihood methods of species tree inference – those that use the observable DNA sequence data directly rather than data summaries such as the collection of inferred gene tree topologies.

2 The Multispecies Coalescent

The multispecies coalescent (MSC) describes the probability distribution of the gene tree, G, underlying a sample of DNA sequences from two or more species (or genetically isolated populations). The MSC is an extension of the coalescent theory for a single randomly mating population. Thus, we begin with a description of the single-population coalescent process.

2.1 The single-population coalescent process

The coalescent theory of population genetics (Kingman, 1982; Hudson, 1983; Tajima, 1983) provides the probability distribution of the gene genealogical history (or gene tree) for a random sample of n sequences at a neutral non-recombining locus. The process is usually formulated in terms of a single parameter

$$\theta = 4N\mu, \tag{1}$$

where N is the effective population size (the population size for an idealized Fisher-Wright model) and μ is the mutation rate per-site per generation. While in classical population genetics, θ is defined using a per-locus mutation rate, the per-site rate used here is far more convenient in analysis of genomic sequence data. Thus θ is the heterozygosity or the average number of mutations per site between two randomly sampled sequences from the population.

The coalescent process tracks the genealogical history of the sequences going backwards in time from the present into the past. The n sequences in the sample go through $n-1$ coalescent events, each time reducing the number of sequences by one, until the most recent common ancestor of the whole sample. With j sequences in the sample, each pair coalesce at the rate $2/\theta$, so that the total rate for $\binom{j}{2}$ pairs is $\binom{j}{2}\frac{2}{\theta}$. The coalescent waiting time until the next coalescence event (which reduces the number of lineages from j to $j-1$) is t_j, is an exponential variable with density

$$f(t_j|\theta) = \binom{j}{2}\frac{2}{\theta}\mathrm{e}^{-\binom{j}{2}\frac{2}{\theta}t_j}, \tag{2}$$

This has expectation $\theta/[j(j-1)]$. The coalescence times $\boldsymbol{t} = \{t_n, t_{n-1}, \cdots, t_2\}$ are independent random variables with joint probability density

$$f(\boldsymbol{t}|\theta) = \prod_{j=2}^{n} f(t_j|\theta) = \prod_{j=2}^{n}\left[\binom{j}{2}\frac{2}{\theta}\mathrm{e}^{-\Sigma_{j=2}^{n}\binom{j}{2}\frac{2}{\theta}t_j}\right], \tag{3}$$

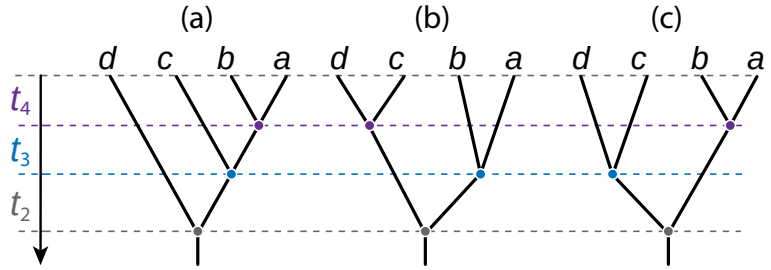


Figure 1 Examples of labelled histories (gene trees with internal nodes rank-ordered according to age) for 4 sequences (a,b,c,d) generated under a coalescent process. There is only one labelled history for a gene tree with the topology $(((a,b),c),d)$ shown in (a) while (b) and (c) are the two alternative labelled histories of the topology $((a,b),(c,d))$ obtained by interchanging the rank order of ages associated with internal nodes.

where time is scaled in units of expected mutations per site. Note that with DNA sequence data, coalescence time (or population size) and mutation rate are not separately identifiable, so that the estimable parameter is $\theta = 4N\mu$, not N and μ separately.

The coalescent process also imposes a probability distribution on gene tree topologies. A "labelled history" (Edwards, 1970) is an ultrametric rooted binary tree with tips labelled and internal nodes rank-ordered according to time or age (see Figure 1). The rank order is completely determined for a fully asymmetrical tree (Figure 1a) but for more symmetrical trees there may be two or more possible rank orderings of the internal nodes (Figure 1b&c). Under the coalescent process all distinct labelled histories have equal probabilities. Degnan and Salter (2005) used a different terminology referring to the alternative orderings of a labelled history as different "instantiations" of the same history. Equation 3 gives the probability density of the times averaged across possible labelled histories. The number of possible labelled histories for n sequences is

$$H_n = \binom{n}{2}\binom{n-1}{2}\cdots\binom{2}{2} = \frac{n!(n-1)!}{2^{n-1}}, \tag{4}$$

Because all labelled histories have equal probability under the process, the probability of the gene tree, $G = \{\boldsymbol{t}, T\}$, defined by a set of coalescence times, $\boldsymbol{t}$, and a labelled history, T, is

$$f(G|\theta) = f(\mathbf{t}|\theta) \times \frac{2^{n-1}}{n!(n-1)!} = \left(\frac{2}{\theta}\right)^{n-1} \mathrm{e}^{-\frac{2}{\theta}\Sigma_{j=2}^{n}\binom{j}{2}t_j}. \tag{5}$$

This probability density applies to a sample from a single panmictic population conforming to a neutral Fisher-Wright model as well as from other neutral models with exchangeable offspring distributions (Kingman, 1982). It can be used within a Bayesian framework for inferring θ using sampled sequences.

The above introduction to the coalescent has focussed on the distribution of the gene trees (topologies and coalescent times) under the model. Many other aspects of the coalescent can be studied as well. Furthermore, the basic neutral coalescent model has been extended to allow for multiple biological processes, including demographic changes over time, recombination (Hudson, 1983; Hudson and Kaplan, 1985; Griffiths and Marjoram, 1996), and selection (Krone and Neuhauser, 1997). The reader may consult Hudson (1990), Nordborg (2007) and Wakeley (2009) for reviews.

2.2 The MSC process

The coalescent process model has been extended to the case of multiple species, which are related through a phylogenetic tree, with one or more sequences sampled from each species. A species

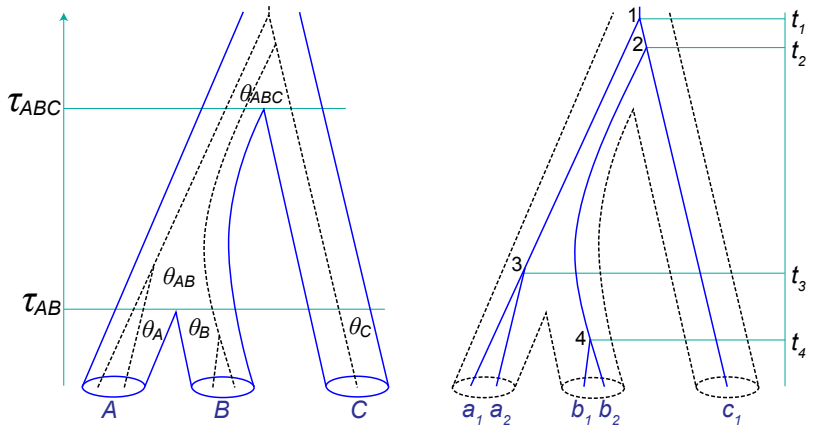


Figure 2 A species tree for three species (A,B,C) with a gene tree for five sequences embedded inside, to illustrate the parameters in the MSC model, $\boldsymbol{\theta} = (\tau_{AB}, \tau_{ABC}, \theta_A, \theta_B, \theta_C, \theta_{AB}, \theta_{ABC})$ and the gene tree density under the model.

tree of s species have $2s-1$ nodes, of which s represent contemporary species and $s-1$ represent ancestral species. The MSC model on a species tree of s species thus has $s-1$ divergence times (τs) and $2s-1$ population size parameters (θs). Both divergence times and population sizes are scaled by mutation rate, so that both τs and θs are measured by expected number of mutations per site. The parameters for a species tree for $s=3$ species are shown in Figure 2. Each population operates as an independent coalescent process during its existence, with population i having a scaled coalescence rate of $\theta_i = 4N_i\mu$. All populations (except the one at the root of the species tree) exist for a finite period of time determined by the species divergence times.

2.2.1 Probability density of gene trees within a species tree

The probability density of an arbitrary gene tree at a locus given the MSC model (species phylogeny with the associated parameters) has been determined by Rannala and Yang (2003). Given the species tree, the gene trees are assumed to be independent among loci. At each locus, the coalescent process is independent among populations on the species tree. Thus we focus on the part of the gene tree residing in one population, say, species X, with parental species P. Let τ_X and τ_P be the age of the two nodes in the species tree. X may be a contemporary species (in which case $\tau_X = 0$) or an ancestral species. Going backwards in time, let m be the number of sequences that enter population X at time τ_X and let $n \geq 1$ be the number of sequences that remain at the end of the population at time τ_P. For example, in figure 2 the species AB (with age τ_{AB}) has parental species ABC (with age τ_{ABC}). In the gene tree in figure 2, $m=3$ lineages enter species AB and $n=2$ lineages leave it. The probability density for the $m-n$ coalescent waiting times between coalescence events is

$$\prod_{j=n+1}^{m}\left[\frac{2}{\theta}\exp\left\{-\frac{j(j-1)}{2}\frac{2}{\theta}t_j\right\}\right] = \left(\frac{2}{\theta}\right)^{m-n}\exp\left\{-\sum_{j=n+1}^{m}\frac{j(j-1)}{\theta}t_j\right\}. \quad (6)$$

An important difference of the MSC from the single-population coalescent is that it is possible for $n \geq 1$ lineages to remain at the end of the population at time τ_P. We have to account for the probability that the n sequences do not coalesce in the remaining period of population existence which has

duration $\left(\tau_P - \tau_X - \sum_{j=n+1}^{m} t_j\right)$. This probability of no events is

$$\exp\left\{-\frac{n(n-1)}{\theta}\left(\tau_P - \tau_X - \sum_{j=n+1}^{m} t_j\right)\right\}. \tag{7}$$

If population X is the root of the species tree, then n must be 1 and this term disappears. Combining the two components gives the probability density for the part of the gene tree in population X

$$\left(\frac{2}{\theta}\right)^{m-n} \exp\left\{-\sum_{j=n+1}^{m}\left(\frac{j(j-1)}{\theta}t_j\right) - \frac{n(n-1)}{\theta}\left(\tau_P - \tau_X - \sum_{j=n+1}^{m} t_j\right)\right\}. \tag{8}$$

The probability density for the whole gene tree at the locus is the product of the probabilities across all populations on the species phylogeny.

For example, given the MSC model for three species in figure 2, the gene tree for the five sampled sequences has the density

$$\begin{aligned} f(G|\boldsymbol{\theta}) &= \left[e^{-\frac{2}{\theta_A}\tau_{AB}}\right] \times \left[\frac{2}{\theta_B}e^{-\frac{2}{\theta_B}t_4}\right] \times \left[\frac{2}{\theta_{AB}}e^{-\frac{3\times 2}{\theta_{AB}}(t_3-\tau_{AB})} \cdot e^{-\frac{2}{\theta_{AB}}(\tau_{ABC}-t_3)}\right] \\ &\times \left[\frac{2}{\theta_{AB}} \cdot \frac{2}{\theta_{AB}}e^{-\frac{3\times 2}{\theta_{ABC}}(t_2-\tau_{ABC})} \cdot e^{-\frac{2}{\theta_{ABC}}(t_1-t_2)}\right]. \end{aligned} \tag{9}$$

The terms in the four pairs of brackets correspond to four species A, B, AB, and ABC, respectively. There is no possibility for coalescent in species C when only one sequence is sampled from the species.

With multiple loci in the data, the probability density for all gene trees is a product over the loci. The formulation allows different sampling configurations at different loci; for example, the number of sequences for each species may vary among loci.

The coalescent is a fundamental process that is operating regardless of whether the species are recently divergent or distantly related, and whether or not the species arose through rapid speciation events so that incomplete lineage sorting is commonplace (Edwards et al., 2016; Degnan, 2018). In cases where species divergences are far apart relative to population sizes, the species tree will have long internal branches and there will be little ILS or gene tree-species tree discordance, but this is exactly as predicted by the MSC model. As discussed by Degnan (2018), the MSC should be considered a null model, and other biological processes, such as recombination, population structure, gene flow, etc. may be incorporated in the model in addition, leading to models such as MSC with recombination, MSC with demographic changes, MSC with migration (which is the IM model Hey, 2010; Hey et al., 2018), MSC with introgression (Yu et al., 2014; Zhang et al., 2018; Wen and Nakhleh, 2018), and so on. Many of these models are not yet implemented because of their complexity, but conceptually they should be possible.

2.2.2 Probabilities of gene tree topologies

Another aspect of the MSC that has been of interest is the marginal probabilities of gene tree topologies conditioned on a particular species tree and branch lengths and, in particular, the probability that the gene tree topology matches that of the species tree (Pamilo and Nei, 1988; Rosenberg, 2002; Degnan and Rosenberg, 2009). As noted above, the labelled histories have equal probabilities under the single population coalescent process. However, this is not the case for the MSC.

The simplest case concerns three species A, B, and C, with three sequences (a, b, c), with one sequence sampled from each species. The probabilities of the three rooted gene tree topologies $G_1 = ((a,b),c)$, $G_2 = ((c,a),b)$ and $G_3 = ((b,c),a)$, given the species tree $S = ((A,B),C)$, were

derived by Hudson (1983). Let the species tree be $((A,B),C)$, the divergence times be τ_{AB} and τ_{ABC}, and the ancestral population size parameters be θ_{AB} and θ_{ABC}. The probability that sequences a and b coalesce in the ancestral population AB (in which case the gene tree must be G_1) is $1-e^{-x} = 1-e^{-(\tau_{ABC}-\tau_{AB})/(\frac{\theta_{AB}}{2})}$, and the probability that sequences a and b do not coalesce in population AB (in which case all three sequences enter the ancestor ABC and the three gene trees occur with equal probability) is e^{-x}. Here $x = 2(\tau_{ABC}-\tau_{AB})/\theta_{AB}$ is known as the internal branch length in coalescent units: one coalescent unit in population AB is $2N_{AB}$ generations or $\theta_{AB}/2$ mutations per site. Thus the probabilities for the three gene tree topologies are

$$P(G_1|S) = (1-e^{-x}) + \tfrac{1}{3}e^{-x}, \tag{10}$$

$$P(G_2|S) = P(G_3|S) = \tfrac{1}{3}e^{-x}. \tag{11}$$

The probabilities that the gene tree matches (or mismatches) the species tree are then

$$P_{\text{match}} = P(G_1|S) = 1-\frac{2}{3}e^{-x}, \tag{12}$$

$$P_{\text{mismatch}} = P(G_2|S)+P(G_3|S) = \frac{2}{3}e^{-x}. \tag{13}$$

In the limit as $x \to \infty$, the probabilities $P_{\text{match}} \to 1$ and $P_{\text{mismatch}} \to 0$ while as $x \to 0$, $P_{\text{match}} \to 1/3$ and $P_{\text{mismatch}} \to 2/3$. Thus, the most difficult species trees to infer using gene trees are those with short internal branches.

Degnan and Salter (2005) developed algorithms for calculating the gene tree probabilities given an arbitrary number of species and arbitrary species tree, with one sequence sampled from each species. The algorithms are computationally expensive owing to the explosive growth in the number of tree topologies with increasing numbers of species and sequences. The gene tree probabilities can be used to estimate the species tree by maximum likelihood, treating the gene tree topologies as data (Wu, 2012, 2016). These are the so-called two-step methods of species tree inference. In practice, almost all two-step methods are based on triplets or quartets, using rooted trees for three species or unrooted trees for four species, and then assembling the results to produce a species tree estimate for all species.

2.2.3 The anomaly zone

The most well-known result from the calculations of gene tree probabilities is the existence of so-called anomaly zone, defined as the zone of species tree and parameter values under which the most probable gene tree has a topology different from the species tree topology (Degnan and Rosenberg, 2006). There is no anomaly zone for three species, but anomaly zone may exist for rooted species trees of four or more species. In the anomaly zone, a "majority-vote" method that uses the most frequent gene tree as the species tree estimate will be inconsistent. Such gene trees are called anomalous gene trees. The anomaly zone exists because the coalescent process generates a uniform distribution on labelled histories but not on rooted tree topologies. As illustrated in Figure 1 asymmetrical topologies have only one labelled history whereas symmetrical topologies can have two or more. This may result in an even greater probability for symmetrical gene-tree topologies even if the species tree has an asymmetrical topology.

Consider the case of four species, related through the asymmetrical phylogeny $(((A,B),C),D)$, and the gene trees for four sequences (a,b,c,d), with one sequence from each species (figure 3). Let the internal branch lengths in coalescent units in the species tree be $x = 2(\tau_{ABCD}-\tau_{ABC})/\theta_{ABC}$ and $y = 2(\tau_{ABC}-\tau_{AB})/\theta_{AB}$. Consider the limit as $x \to 0$ and $y \to 0$. In this case, the probability of a coalescence in either ancestral species AB or ABC approaches zero and all coalescence events will occur in the root species $ABCD$. The coalescent process in $ABCD$ is equivalent to a single

population coalescent with four sequences so applying Equation 4 we have 18 possible labelled histories, each with equal probability $1/18$. Of these, 12 are fully asymmetrical (such as labelled histories G_2 and G_3 in figure 3) and 6 are symmetrical (such as labelled history G_1 in figure 3). Each asymmetrical labelled history corresponds to one unique tree topology because there is only one possible way to order the internal nodes. The 6 symmetrical labelled histories form 3 pairs, with each pair corresponding to one tree topology with two possible node orderings (e.g., G_2 and G_3 in figure 3). Thus, each symmetrical rooted tree topology receives probability $1/18+1/18=2/18$ whereas each asymmetrical rooted tree topology receives probability $1/18$. When the branch lengths x and y are nonzero but very small, the symmetrical and mismatching gene tree (corresponding to G_2 and G_3 in figure 3) may still be more frequent than the asymmetrical and matching gene tree G_1, even if not twice as frequent. Consequently, the most common gene tree will have a mismatching symmetrical topology that is different from the species tree, and this combination of species tree topology and branch lengths is in the anomaly zone.

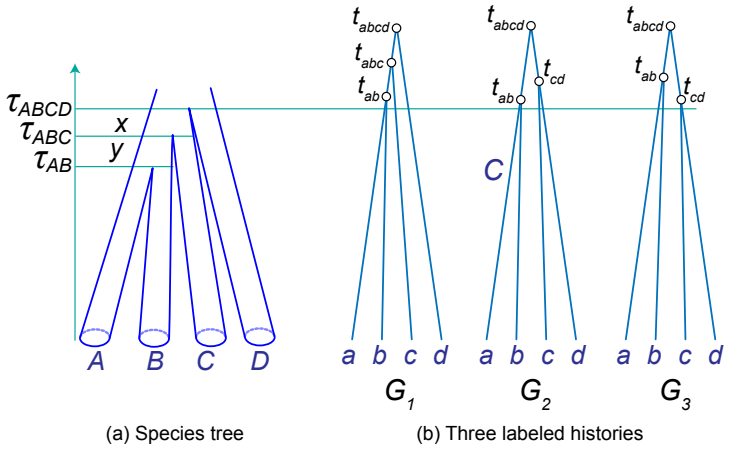


Figure 3 A species tree for four species (A,B,C,D) with very short internal branches and three labelled histories for four sequences (a,b,c,d) to illustrate the existence of the anomaly zone.

The anomaly zone has been shown to affect empirical dataset from lizards (Linkem et al., 2016), flightless birds (Cloutier et al., 2019), gibbons (Shi and Yang, 2018), and African mosquitoes (Thawornwattana et al., 2018). The anomaly zone can be identified by estimating parameters in the MSC model using Bayesian inference programs such as bpp, and then simulating gene trees using those parameters to estimate gene tree probabilities –to confirm that the most probable gene tree does not match the species tree (Shi and Yang, 2018).

While it is well-known that species phylogenies with very short internal branches are hard to recover, the importance of the anomaly zone may have been exaggerated in the literature. Note that the anomaly zone is the zone of inconsistency for the simple "majority vote" method only. Other methods may, or may not, be inconsistent in the anomaly zone. In particular, methods based on the likelihood function for the sequence data, including maximum likelihood and Bayesian methods (see below), are consistent both inside and outside the anomaly zone; indeed they are consistent over the entire space of species trees (Xu and Yang, 2016).

3 Species Tree Inference Methods

The MSC provides a framework for developing parametric multi-locus statistical methods for species tree inference. Such methods allow gene trees to differ from species trees due to ILS and provide

estimates of ancestral demographic parameters. Because the MSC operates in all finite populations it is the canonical model for species tree inference. We begin by describing the maximum likelihood and Bayesian methods that have been developed for species tree inference. These are often referred to as full-likelihood methods because they use an exact likelihood function. Full-likelihood methods are known to possess optimal statistical properties such as consistency and efficiency. We then consider two of the most widely used approximate methods: MP-EST (Liu et al., 2010) and ASTRAL (Mirarab et al., 2014; Mirarab and Warnow, 2015). These programs are examples of "super-tree" methods which infer larger trees by combining estimates of smaller trees. One of the methods (MP-EST) approximates the MSC using pseudo-likelihood while the other (ASTRAL) uses a simple heuristic that may provide estimates that are statistically consistent when gene trees arise under the MSC. The statistical properties of heuristic methods often can only be studied by computer simulation. See Yang and Rannala (2014), Edwards (2016) and Xu and Yang (2016) for an overview of other approximate methods.

Next, we consider another class of approximate methods, so-called concatenation methods. These methods combine all the loci into a single matrix of sequences and are examples of "super-gene" or "super-matrix" approaches to species tree inference that implicitly assume no ILS. Gatesy and Springer (2013) and Edwards et al. (2016) review the extensive discussions concerning relative strengths and weaknesses of concatenation versus two-step and coalescent methods for species tree inference. We briefly summarize several of the problems that can arise with approximate inference methods that use concatenation. The reader may consult Chapter 3.4 (Bryant and Hahn 2020) for a different perspective. Finally, we discuss some criticisms of two-step approximate inference methods and full-likelihood methods based on the MSC.

3.1 Maximum likelihood method

Here we outline full-likelihood methods for estimating the species tree using multilocus sequence data under the MSC. Let the sequence alignment at locus i be X_i, with $i = 1, 2, \cdots, L$. Let $\boldsymbol{X} = \{X_i\}$. Let $\boldsymbol{\theta}_k$ be the MSC parameters (θs and τs) in species tree S_k. Let G_i be the gene tree at locus i. The main difference from traditional phylogenetic methods is that the gene trees are unobserved random variables, with distributions specified by the MSC model (Rannala and Yang, 2003). For example, the maximum likelihood method of species tree estimation maximizes the following likelihood function

$$f(\boldsymbol{X}|S_k, \boldsymbol{\theta}_k) = \prod_{i=1}^{L} \left[\int f(G_i|S_k, \boldsymbol{\theta}_k) f(X_i|G_i) \, \mathrm{d}G_i \right], \tag{14}$$

where $f(G_i|S_k, \boldsymbol{\theta}_k)$ is the MSC density for gene tree G_i at locus i discussed above (Rannala and Yang, 2003), and $f(X_i|G_i)$ is the probability of the sequence alignment at locus i or the phylogenetic likelihood (Felsenstein, 1981). The integral over gene tree G_i represents a summation over all possible gene tree topologies (labelled histories) for the locus and an integral over the coalescent times within each gene tree topology. In this formulation, the gene trees G_i are unobserved random variables (called latent variables), and the likelihood function for the species tree and MSC parameters has to average over all possible gene trees at each locus.

The S_k and $\boldsymbol{\theta}_k$ that maximize the log-likelihood, $\ell = \log f(\boldsymbol{X}|S_k, \boldsymbol{\theta}_k)$, will be the ML species tree and MLEs of parameters in that species tree. Note that both the MSC density $f(G_i|S_k, \boldsymbol{\theta}_k)$ and the phylogenetic likelihood $f(X_i|G_i)$ are straightforward to calculate. The difficulty with the ML method lies in the averaging over the possible gene trees at each locus, because the number of possible gene trees is huge and the integral over coalescent times for each gene tree at a locus with n sequences is $(n-1)$-dimensional. The only ML implementation that has been achieved is the 3s program (Yang, 2002; Dalquen et al., 2017), which enumerates the gene tree topologies and uses numerical integration

(Gaussian quadradutre) to calculate the integrals. This is limited to three species and three sequences per locus, but can accommodate tens of thousands of loci.

3.2 Bayesian inference

In a Bayesian approach, we specify a prior probability distribution for all possible species trees, and for each species tree (which is an MSC model), we specify a prior for the parameters of the MSC model (θs and τs). Let $f(S_k)$ be the prior probability for species tree S_k. This can be a uniform distribution on the rooted species trees or on the labelled histories –ranked trees (Yang and Rannala, 2014). It is common to assign gamma or inverse priors on the MSC parameters given the species tree, $f(\boldsymbol{\theta}_k|S_k)$. Inverse gamma priors are conjugate on θ, allowing θs to be integrated out analytically (Hey and Nielsen, 2007). This may improve the Markov chain Monte Carlo (MCMC) mixing slightly during the tree search thanks to the reduced parameter space. Usually a gamma or inverse-gamma prior is assigned on the age of the root (τ_0), while the other node ages may be specified by a Dirichlet distribution (Yang and Rannala, 2010) or by a birth-death process model. Bayesian computation is achieved through MCMC algorithms, which generate a sample from the joint conditional distribution (joint posterior) of the species tree and the gene trees

$$f(S_k, \boldsymbol{\theta}_k, \{G_i\}|\boldsymbol{X}) \propto f(S_k) f(\boldsymbol{\theta}_k|S_k) \prod_{i=1}^{L} [f(G_i|S_k, \boldsymbol{\theta}_k) f(X_i|G_i)]\,. \tag{15}$$

Compared with Equation 14, the integral over the gene trees disappears; instead integration occurs numerically through the MCMC. In other words, MCMC is used to traverse the joint space of gene trees as well as the species tree and the MSC parameters. The frequency at which the MCMC visits each species tree is the estimate of the posterior probability for that species tree. The first implementation of the Bayesian method is the Best program (Liu and Pearl, 2007), which used the posterior sample of gene trees from the MrBayes program (Ronquist and Huelsenbeck, 2003) and applied a correction for the gene tree density, because the gene trees from MrBayes are not generated with an MSC prior. Later implementations work directly on the MSC and sequence alignments, rather than processing MrBayes outputs (Liu, 2008), especially in the programs StarBeast (Heled and Drummond, 2010; Ogilvie et al., 2017) and bpp (Yang and Rannala, 2014; Rannala and Yang, 2017). Branch-swapping algorithms in phylogenetic tree search such as nearest-neighbor-interchange (NNI), subtree-pruning-and-regrafting (SPR), etc. have been adapted to become MCMC proposals, proposing changes from one species tree to another (Yang and Rannala, 2014; Rannala and Yang, 2017; Flouri et al., 2018), while other moves are used to change the gene trees.

The greatest challenge for MCMC algorithms for species tree inference appears to be the constraint between the species tree and the gene trees. If the species tree is changed when the gene trees at all loci are fixed, the current gene trees may place extremely stringent constraints on the species tree. Consider the simple move that changes the divergence time τ_{AB} between two species or clades A and B. In the coalescent model, the sequence divergence has to be older than species divergence, that is, $t_{ab} > \tau_{AB}$, and this constraint applies to every pair of sequences from A and B at every locus. In other words current gene trees provide a maximum bound for τ_{AB}, which is the minimum t_{ab} across all loci. If there are thousands of loci in the dataset and many sequences from A and B at each locus, the current value of τ_{AB} is often almost identical to this bound and τ_{AB} cannot possibly become even greater, and the algorithm is essentially stuck. An algorithm that appears to work well is the rubber-bound algorithm implemented in bpp (Rannala and Yang, 2003), which identifies the nodes on the gene trees that are affected by the change of τ_{AB}, and then modifies the ages of those gene-tree nodes at the same time that τ_{AB} is changed, in the same way that marked points on a rubber band move when with its two ends fixed, a rubber bound is pulled from a point in the middle to one end. This move has been ported into StarBeast (Jones, 2017). Similar coordinated changes between the

species tree and the gene trees appear to improve MCMC mixing when the species tree topology is changed (Yang and Rannala, 2014; Rannala and Yang, 2017; Jones, 2017; Ogilvie et al., 2017). Those smart MCMC moves have made it possible to analyze large datasets with more than 10,000 loci (Rannala and Yang, 2017; Shi and Yang, 2018; Thawornwattana et al., 2018). Further improvements in both computational and mixing efficiency are clearly needed, as real datasets are often too large for Bayesian MCMC programs to handle.

3.3 Approximate species tree inference methods

Next-generation sequencing technologies are currently advancing at an astounding rate making dense genome sequences available for hundreds of individuals and species. This vast array of new data has driven demand for computational methods for inferring species trees that can be practically applied with thousands of loci and hundreds or thousands of sequences. Model-based methods for species tree estimation, such as the Bayesian inference procedure described above, are computationally intensive, and will lag behind the demands of many contemporary sequencing projects. As a result, many heuristic (or approximate) species tree inference methods have been proposed that use various shortcuts and heuristic approximations to improve computational efficiency for large datasets. Some of the heuristic methods discussed (MP-EST and ASTRAL) make an explicit attempt to accommodate ILS, while others (concatenation methods) do not.

One class of approximate species tree inference methods (super-tree methods such as MP-EST and ASTRAL) take the two-step approach of estimating the gene trees from phylogenetic analysis of sequence alignments at individual loci and then treating the gene trees as observed data. A second class of approximate species tree inference methods (sometimes referred to as supermatrix or super-gene methods) concatenate all the loci into a single sequence assuming that the gene tree of the supermatrix matches the species tree. This approach typically applies a standard maximum likelihood or Bayesian phylogenetic approach under the assumption of one gene tree that matches the species tree.

Two-step super-tree methods are much simpler to implement than full likelihood methods and are among the early approaches developed for inferring species trees under the MSC. Some of them use the estimated gene tree topologies with branch lengths (node ages), such as the Maximum Tree method of (Liu et al., 2010) implemented in the STEM program (Kubatko et al., 2009). A serious problem is that the method does not account for the sampling errors in the estimated tree topology and coalescent times. In particular, the coalescent times can have a major impact on species tree inference: for example, if two sequences from two species or clades A and B are identical at any locus, with $t_{ab} = 0$, then the species divergence time must be $\tau_{AB} = 0$. Such extreme estimates of species divergence times will influence the inference of the species tree topology. Other two-step methods use the estimated gene tree topologies as data, ignoring branch lengths or coalescent times. These methods use less information from the data but are also less affected by phylogenetic reconstruction errors. They often work on unrooted gene trees, which are estimated without the assumption of the molecular clock. These topology-only methods have been more successful than methods based on inferred gene trees with branch lengths. However, many of the two-step methods have poor statistical performance and the accuracy of some methods (such as the two-step likelihood method STEM [Kubatko et al. 2009]) even decreases with increasing numbers of loci (Leaché and Rannala, 2010; Mirarab et al., 2014). Concatenation-based super-gene methods rely on straightforward application of existing single-locus phylogenetic inference methods and are thus simple to apply. However, differences between gene trees and species trees (both in terms of branch lengths and topologies) resulting from the MSC and other processes can cause the methods to be statistically inconsistent.

3.3.1 MP-EST: Maximum Pseudo-likelihood Estimation

The Maximum Pseudo-likelihood Estimation (MP-EST) method (Liu et al., 2010) is a two-step method based on species triplets. It extracts, for a tree with s species, all the $s(s-1)(s-2)/6$ rooted triplet "species subtrees" (each comprised of 3 species) to construct the likelihood function. The single internal branch length in each rooted species subtree is a sum of one or more internal branch lengths in the original s-species tree. The data input to the program are rooted gene tree topologies inferred using a maximum likelihood or Bayesian inference program. The number of each triplet gene tree topology given the species subtree follows a trinomial distribution with probabilities determined by the MSC (Equation 10)). The probabilities for gene tree topologies for triplets are multiplied across species subtrees and across loci. This is a pseudo-likelihood function as it ignores the fact that the triplet subtrees are not independent. The pseudo-likelihood is maximized using a heuristic search algorithm to infer the species tree. The MP-EST method may suffer from an information loss because it ignores branch lengths in the gene trees and because it ignores phylogenetic errors in the gene tree reconstruction; this is true of all super-tree methods.

Note that the theory underlying the MP-EST method is given in Equation 10. With the probabilities for the three gene trees given, one can find the most common gene tree topology, which is the species tree estimate, and estimate the internal branch length in the species tree (x). The method is clearly consistent, if the gene trees are known without error: when the number of loci or gene trees approaches infinity, the probability of recovering the correct species tree topology approaches one. Furthermore, in this case of three species and rooted gene trees, Yang (2002) showed that the most probable *estimated* gene tree topology is the one that matches the species tree, although phylogenetic reconstruction errors have the effect of inflating the gene tree-species tree mismatch probability. Thus the MP-EST method will be consistent when *estimated* gene tree topologies are used to estimate the species tree. The internal branch length in the species tree, however, is inconsistently estimated (and underestimated) because phylogenetic errors distort the gene tree probabilities and inflate the gene tree-species tree discordance.

3.3.2 ASTRAL: Accurate Species Tree Algorithm

ASTRAL (Mirarab et al., 2014; Mirarab and Warnow, 2015) is another two-step program that takes as input unrooted gene trees inferred using the maximum likelihood phylogenetic program RAxML (Stamatakis, 2006). The underlying method is based on quartets, with four species and four sequences, one sequence sampled from each species. The species tree is then chosen to be the one that agrees with the greatest number of quartet gene trees. If multiple sequences are available from one species, one sequence from each species is sampled to form the quartet. A motivation for using unrooted quartets for the optimization, rather than finding the species tree compatible with the largest number of complete gene trees (the "majority-vote tree") is that there are no anomalous gene trees in the case of unrooted species trees for four species (Degnan, 2013).

The ASTRAL method essentially uses ML estimates of the gene tree topologies as summary statistics for inference. It does not use branch length information from the gene trees. Use of the gene tree topologies alone allows the identification of the species tree topology, as well as the internal branch lengths in coalescent units, but other parameters in the MSC model are not identifiable. Note that while the method is claimed to be consistent, the proof of consistency relies on the assumption that gene trees are known without error, and the impact of phylogenetic reconstruction errors is, in general, unknown although this is sometimes evaluated using computer simulation (Huang and Knowles, 2009).

3.3.3 Concatenation methods

A simple approach to inferring the species tree using multi-loci sequence data is to concatenate the sequences across loci and then infer a single tree using the "super-gene" sequence as the species tree estimate. This implicitly assumes that all gene loci share the same topology and branch lengths. Systematists have long struggled with the issue of whether to combine different genes into a single analysis (de Queiroz et al., 1995). From a statistical viewpoint, a standard approach for analyzing heterogeneous data is to do a combined analysis accommodating heterogeneity (Yang, 1996). However, until the development and implementation of the MSC model, no formal statistical method existed allowing multiple genes to be combined while respecting their different histories. When the species tree is easy, with long internal branches and small population sizes, one expects very little deep coalescence or incomplete lineage sorting. In such cases, concatenation and coalescent methods are expected to yield the same species topology (Edwards et al., 2007; Leaché and Rannala, 2010; Kubatko and Degnan, 2007). However, when the species tree is challenging, with short internal branches and large population sizes, concatenation may be inconsistent and may converge to an incorrect species tree topology (Roch and Steel, 2015).

Even if gene trees share topology, they may have different branches (coalescent times) due to coalescent fluctuations. In such cases, concatenation can lead to biases in estimation of major evolutionary parameters such as species divergence times, while coalescent methods (full-likelihood methods applied to sequence alignments) accommodate variable coalescence times providing reliable estimates (Ogilvie et al., 2017). A recent Bayesian analysis of diverse phylogenomic data sets (Jiang et al., 2019) suggests that (i) gene tree heterogeneity is real and abundant, even after accounting for gene tree errors; (ii) the concatenation assumption of topologically congruent gene trees can be rejected in almost all datasets; and iii) the MSC model fits phylogenomic datasets better than the concatenation model. Concatenation continues to be a widely used approach (see Chapter 2.1 [Simion et al. 2020]), especially in comparative analyses of recently sequenced genomes, mainly because of its simplicity and lower computational burden. With the development of improved algorithms for MSC-based species tree inference and broader recognition of the importance of accommodating the coalescent process within species this situation may change.

3.4 Criticisms of MSC species tree inference methods

MSC-based methods of species tree inference make the assumption of no intra-locus recombination. Gatesy and Springer (2013) correctly noted that when multiple exons in transcriptome data are concatenated into one gene or locus, the exons may span large distances along the chromosome; this hybrid concatenation-coalescence approach may lead to violation of the MSC model. Based on empirical calculations, Springer and Gatesy (2016) predicted that the non-recombining unit in a typical species radiation is short enough to violate the MSC assumption of no recombination. However, their calculation does not account for the fact that recombination events during the time period when there is only one sequence in the sample are consistent with the MSC assumption (Edwards et al., 2016). Furthermore, simulation suggests that intra-locus recombination may be a problem for MSC methods under extreme levels of ILS only (Lanier and Knowles, 2012). The assumption of no recombination is more problematic for concatenation than for two-step coalescent methods because concatenation assumes the same genealogical history for all sites in all genes, which is almost certainly violated.

As noted above, two-step coalescent methods treat estimated gene trees as data and do not account for phylogenetic errors; this can cause two-step methods to underestimate internal branches in species trees and exaggerate the importance of ILS by inflating gene tree vs species tree discordance (Yang, 2002; Mirarab et al., 2016; Springer and Gatesy, 2014, 2016). This criticism applies to

"two-step" coalescent methods specifically, because full likelihood methods accommodate gene tree errors correctly through the phylogenetic likelihood function (Equations 14 and 15). Recent efforts making use of the bootstrap and other measures of gene-tree uncertainty to correct for phylogenetic uncertainties in two-step methods may help reduce the impact of phylogenetic errors (Sayyari and Mirarab, 2016). The above discussion largely applies to shallow phylogenies for closely related species. For deep phylogenies involving distantly related species, a whole suite of complicating factors that affect phylogenetic analysis will affect species tree inference as well, including violation of the molecular clock, heterogeneity in the substitution process across genomic loci and across lineages (Yang, 2014). These factors operate in addition to deep coalescence, making inference of deep phylogenies for species that arose through ancient radiative speciation events a very challenging task (see Chapter 3.4 [Bryant and Hahn 2020]). We note that model violation is a common feature in phylogenetics, and whether a misspecified model is still useful may depend on a number of factors including the impact of the model on the analysis (see Chapter 2.1 [Simion et al. 2020]). More complex models, especially MSC models that account for migration or introgression, are likely to be even better than the basic MSC without gene flow and may lead to improved inference under complex scenarios where both deep coalescence and introgression exist (Bravo et al., 2019; Edwards et al., 2016; Nakhleh, 2013; Yu et al., 2013; Zhang et al., 2018).

4 Future Challenges

The development of the multispecies coalescent model is a major advance in molecular phylogenetics (Edwards, 2009). The model accommodates fluctuations in genealogical history across the genome and provides a natural framework for inference using genomic sequence data from closely related species, bridging the gap between phylogenetics and population genetics. The MSC forms the basis for addressing many exciting inference problems in phylogenomics and population genomics, including estimation of ancestral population sizes and inference of ancient hybridisation events – even those hybridization events involving species that have since gone extinct (Xu and Yang, 2016; Degnan, 2018).

Currently full-likelihood implementations of the MSC model, mostly in the form of MCMC algorithms, involve intensive computation. With the increase of data size (e.g., the number of species, the number of sites per sequence, the number of sequences per locus, and the number of loci), each MCMC iteration takes more computational effort. Furthermore there is a deterioration in MCMC mixing so that more MCMC iterations are necessary to generate an acceptable effective sample size. Most of the current MCMC implementations are not computationally feasible for genome-scale datasets with thousands of loci (see Chapter 1.4 [Lartillot 2020]), although implementations of smart MCMC moves in bpp that propose coordinated changes to both the gene trees and the species tree have made it possible to analyse datasets with over 10,000 loci (Rannala and Yang, 2017; Flouri et al., 2018). Further improvements in the computational and mixing efficiency of the algorithms are highly desirable.

The explosive growth of genomic sequence data means that approximate or heuristic methods will continue to play a major role in data analysis (see Chapter 1.2 [Stamatakis and Kozlov 2020]). Current two-step methods appear to make use of only a small portion of the information in genomic datasets, in particular in analysis of shallow phylogenies for closely related species, and as a result many parameters in the MSC model are unidentifiable by the two-step methods, even though the species tree topology is. Development of statistically more efficient heuristic methods should be a priority for future research.

For distantly related species, the molecular clock may be seriously violated. Even though one can adapt the relaxed-clock models developed in phylogenetics (dos Reis et al., 2016) to accommodate

the violation of the clock, the rate variation means that some of the temporal information in gene trees is eroded. It remains to be seen how full likelihood methods under the MSC with relaxed clock compare with heuristic methods using unrooted gene tree topologies and ignoring time information in gene-tree branch lengths.

We expect that accommodating cross-species gene flow in the MSC model will be a research hotspot in the next few years. Many recent empirical studies suggest that cross-species gene flow may be commonplace in animals as well as plants and indeed across the tree of life (Mallet et al., 2016; Folk et al., 2018; Degnan, 2018). The MSC model can be extended to accommodate cross-species gene flow. Two such models have been developed. The MSC-with-migration model, better known as the isolation-with-migration (IM) model (Hey and Nielsen, 2004), assumes continuous migration, with species exchanging migrants at certain rates every generation. This model is similar to population genetics models of population subdivision except that under the IM model the populations have a phylogenetic history with a branching order and divergence times. The probability density of the gene trees under the IM model is given by Hey and Nielsen (2004) and Hey (2010). The MSC with introgression (MSci) model (Flouri et al., 2020), also known as multispecies network coalescent (MSNC) model (Wen and Nakhleh, 2018), assumes episodic introgression/hybridization; in other words, introgression happened at a certain time point in the past. Important parameters in the model include the time of introgression and introgression probability. The gene tree density under the MSci model is given by Yu et al. (2014). Bayesian MCMC implementations include IMa3 (Hey, 2010; Hey et al., 2018) for the IM model, and StarBeast (Zhang et al., 2018; Jones, 2019) and PhyloNet (Wen and Nakhleh, 2018) for the MSci model. Those programs involve expensive computation and are not feasible for realistically sized datasets, with more than 200 loci, say. At the same time, the complexity of those models means that large datasets with thousands of loci may be necessary to obtain reliable parameter estimates. In the case of the IM model, the MCMC averages over a huge space of genealogical history at each locus, which includes the number and directions of migration events. At high migration rates, this space is in effect infinite and the likelihood surface is nearly flat over this space, because the sequence likelihood depends on the gene tree topology and divergence times but not on migration events. For the MSci model, a major stumbling block is the constraint between the species tree or network and the gene trees, as in the case of the simple MSC model. A recent effort to develop coordinated moves between the model parameters such as species divergence or hybridisation times and the gene trees in bpp has made it possible to analyse data of more than 10,000 loci (Flouri et al., 2020), but currently the model is fixed, with the number and directions of the introgression events specified by the user. MCMC proposals to allow moves between different MSci models are yet to be implemented. There is an urgent need to improve the computational efficiency of the full likelihood methods.

Several heuristic methods have been developed to detect cross-species gene flow and to estimate the introgression probability. Some take the two-step approach and use estimated gene tree topologies, such as SNaQ (Solis-Lemus et al., 2016, 2017). Others use other summaries of the multi-locus sequence data such as the counts of parsimony-informative site patterns for three or four species, including the popular ABBA-BABA test (Green et al., 2010; Durand et al., 2011) and the HyDe program (Blischak et al., 2018). Those methods do not use information in branch lengths on gene trees, although the recent heuristic method of Hibbins and Hahn (2019) does attempt to use branch length information. They can estimate the introgression probability and internal branch lengths in coalescent units on the species tree but other parameters in the model are unidentifiable. Moreover, many introgression scenarios are not identifiable and cannot be detected using those methods. In cases where the introgression parameter is identifiable the two-step methods appear to provide estimates with similar accuracy to full likelihood methods (Flouri et al., 2020). Developing statistically efficient heuristic methods should be a high priority in the next few years.

Another important avenue for future research concerning the MSci models is their identifiability (Degnan, 2018). The data may be either gene tree topologies (for the two-step heuristic methods) or multilocus sequence alignments (for full likelihood methods). Identifiability may concern either different introgression models (which assume different numbers of introgression events or assume introgressions involving different species) or parameters in a given introgression model (including the species divergence times, population sizes, and introgression probabilities). Some of the identifiability issues might be solved by using more informative summary statistics in two-step methods but that will likely make the derivation of a heuristic estimator more difficult.

Species tree inference is a difficult statistical problem, especially when factors such as introgression are incorporated. The MSC is a model that links population genetics with evolutionary history and it is for this reason central to the problem of species tree inference. We expect that the objective of efficiently and accurately inferring species trees will remain at the heart of the discipline of phylogenetic inference for the foreseeable future. Although much progress has been made during the last two decades many challenging problems remain.

References

Avise, J. C. (1994). *Molecular Markers, Natural History and Evolution*. Chapman and Hall, New York.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., and Saunders, N. C. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, pages 489–522.

Avise, J. C., Lansman, R. A., and Shade, R. O. (1979). The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. i. population structure and evolution in the genus peromyscus. *Genetics*, 92(1):279–95.

Blischak, P. D., Chifman, J., Wolfe, A. D., and Kubatko, L. S. (2018). HyDe: A python package for genome-scale hybridization detection. *Syst. Biol.*, 67(5):821–829.

Boussau, B., Szollosi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Res*, 23(2):323–30.

Bravo, G. A., Antonelli, A., Bacon, C. D., Bartoszek, K., Blom, M. P. K., Huynh, S., Jones, G., Knowles, L. L., Lamichhaney, S., Marcussen, T., Morlon, H., Nakhleh, L. K., Oxelman, B., Pfeil, B., Schliep, A., Wahlberg, N., Werneck, F. P., Wiedenhoeft, J., Willows-Munro, S., and Edwards, S. V. (2019). Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. *PeerJ*, 7:e6399.

Brown, W., Prager, E., and Wilson, A. (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution*, 18:225–39.

Bryant, D. and Hahn, M. W. (2020). The concatenation question. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.4, pages 3.4:1–3.4:23. No commercial publisher | Authors open access book.

Cloutier, A., Sackton, T. B., Grayson, P., Clamp, M., Baker, A. J., and Edwards, S. V. (2019). Whole-genome analyses resolve the phylogeny of flightless birds (palaeognathae) in the presence of an empirical anomaly zone. *Systematic Biology*, 68(6):937–955.

Dalquen, D., Zhu, T., and Yang, Z. (2017). Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst. Biol.*, 66:379–398.

Darwin, C. (1859). *On the Origin of Species*. Harvard University Press, Cambridge.

de Queiroz, A., Donoghue, M. J., and Kim, J. (1995). Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics*, 26:657–681.

Degnan, J. H. (2013). Anomalous unrooted gene trees. *Systematic Biology*, 62(4):574–590.

Degnan, J. H. (2018). Modeling hybridization under the network multispecies coalescent. *Syst. Biol.*, 67(5):786–799.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5):e68.

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.*, 24:332–340.

Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59:24–37.

dos Reis, M., Donoghue, P. C. J., and Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.*, 17:71–80.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Mol. Bio.l Evol.*, 28:2239–2252.

Edwards, A. W. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(2):155–164.

Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, 63(1):1–19.

Edwards, S. V. (2016). Inferring species trees. In Kliman, R., editor, *Encyclopedia of Evolutionary Biology*. Elsevier, New York.

Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14):5936–5941.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.*, 94(Pt A):447–462.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2018). Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, 35(10):2585–2593.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2020). A bayesian implementation of the multispecies coalescent model with introgression for comparative genomic analysis. *Mol. Biol. Evol.*, page under review.

Folk, R. A., Soltis, P. S., Soltis, D. E., and Guralnick, R. (2018). New prospects in the detection and comparative analysis of hybridization in the tree of life. *Am. J. Bot.*, 105(3):364–375.

Gatesy, J. and Springer, M. S. (2013). Concatenation versus coalescence versus "concatalescence". *Proceedings of the National Academy of Sciences of the United States of America*, 110(13):E1179.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Paabo, S. (2010). A draft sequence of the neandertal genome. *Science*, 328:710–722.

Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4):479–502.

Hare, M. (2001). Prospects for nuclear gene phylogeography. *Trends in Ecology and Evolution*, 16:700–706.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27:570–580.

Hey, J. (2010). Isolation with migration models for more than two populations. *Mol. Biol. Evol.*, 27:905–920.

Hey, J., Chung, Y., Sethuraman, A., Lachance, J., Tishkoff, S., Sousa, V. C., and Wang, Y. (2018). Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.*, 35(11):2805–2818.

Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 167:747–760.

Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A*, 104(8):2785–2790.

Hibbins, M. S. and Hahn, M. W. (2019). The timing and direction of introgression under the multispecies network coalescent. *Genetics*, 211(3):1059–1073.

Huang, H. and Knowles, L. L. (2009). What is the danger of the anomaly zone for empirical phylogenetics? *Syst. Biol.*, 58:527–536.

Hudson, R. (1990). Gene genealogies and the coalescent process. In Futuyma, D. and Antonovics, J. D., editors, *Oxford Surveys in Evolutionary Biology*, pages 1–44. Oxford University Press, New York.

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, pages 203–217.

Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–164.

Jiang, X., Edwards, S., and Liu, L. (2019). The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. *bioRxiv*.

Jones, G. (2017). Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *J. Math. Biol.*, 74:447–467.

Jones, G. R. (2019). Divergence estimation in the presence of incomplete lineage sorting and migration. *Syst. Biol.*, 68(1):19–31.

Kingman, J. F. (1982). The coalescent. *Stochastic Processes and Their Applications*, 13(3):235–248.

Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Pääbo, S., Villablanca, F. X., and Wilson, A. C. (1989). Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences (USA)*, 86:6196–6200.

Krone, S. M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology*, 51(3):210–237.

Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009). STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7):971–973.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24.

Lanier, H. C. and Knowles, L. L. (2012). Is recombination a problem for species-tree analyses? *Systematic Biology*, 61(4):691–701.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Leaché, A. D. and Rannala, B. (2010). The accuracy of species tree estimation under simulation: a comparison of methods. *Systematic Biology*, 60(2):126–137.

Linkem, C. W., Minin, V. N., and Leaché, A. D. (2016). Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (squamata: Scincidae). *Systematic Biology*, 65(3):465–477.

Liu, L. (2008). BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21):2542–2543.

Liu, L. and Pearl, D. K. (2007). Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.*, 56(3):504–514.

Liu, L., Xi, Z., Wu, S., Davis, C. C., and Edwards, S. V. (2015). Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences*, 1360:36–53.

Liu, L., Yu, L., Kubatko, L., Pearl, D. K., and Edwards, S. V. (2009). Coalescent methods for estimating phylogenetic trees. *Mol Phylogenet Evol*, 53(1):320–8.

Liu, L., Yu, L., and Pearl, D. K. (2010). Maximum tree: a consistent estimator of the species tree. *J. Math. Biol.*, 60:95–106.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

Mallet, J., Besansky, N., and Hahn, M. W. (2016). How reticulated are species? *BioEssays*, 38(2):140–149.

Mirarab, S., Bayzid, M., and Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65:366–380.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Mirarab, S. and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52.

Nakhleh, L. (2013). Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology and Evolution*, 28(12):719–728.

Nichols, R. (2001). Gene trees and species trees are not the same. *Trends Ecol. Evol.*, 16:358–364.

Nordborg, M. (2007). Coalescent theory. In Balding, D., Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*, pages 843–877. Wiley, San Francisco.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*, 34(8):2101–2114.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5):568–583.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656.

Rannala, B. and Yang, Z. (2008). Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics*, 9:217–231.

Rannala, B. and Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, 66:823–842.

Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.*, 100:56–62.

Ronquist, F. and Huelsenbeck, J. P. (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574.

Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology*, 61(2):225–247.

Sayyari, E. and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.*, 33(7):1654–1668.

Shi, C. and Yang, Z. (2018). Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.*, 35:159–179.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Solis-Lemus, C., Bastide, P., and Ane, C. (2017). PhyloNetworks: A package for phylogenetic networks. *Mol. Biol. Evol.*, 34(12):3292–3298.

Solis-Lemus, C., Yang, M., and Ane, C. (2016). Inconsistency of species tree methods under gene flow. *Syst. Biol.*, 65(5):843–851.

Springer, M. S. and Gatesy, J. (2014). Land plant origins and coalescence confusion. *Trends in Plant Science*, 19(5):267–9.

Springer, M. S. and Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94:1–33.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Szollosi, G. J., Tannier, E., Daubin, V., and Boussau, B. (2015). The inference of gene trees with species trees. *Syst. Biol.*, 64(1):e42–62.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.

Takahata, N., Satta, Y., and Klein, J. (1995). Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.*, 48:198–221.

Thawornwattana, Y., Dalquen, D., and Yang, Z. (2018). Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Mol. Biol. Evol.*, 35(10):2512–2527.

Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Roberts & Company, Greenwood Village, Colorado.

Wen, D. and Nakhleh, L. (2018). Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst. Biol.*, 67(3):439–457.

Wilson, A., Cann, R. L., Carr, S. M., George, M., Gyllensten, U. B., Helm-Bychowski, K. M., Higuchi, R. G., Palumbi, S. R., Prager, E. M., Sage, R. D., and Stoneking, M. (1985). Mitochondrial DNA and two perspectives on evolutionary genetics. *Biological Journal of the Linnaean Society*, 26:375–400.

Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution: International Journal of Organic Evolution*, 66(3):763–775.

Wu, Y. (2016). An algorithm for computing the gene tree probability under the multispecies coalescent and its application in the inference of population tree. *Bioinformatics*, 32(12):i225–i233.

Xu, B. and Yang, Z. (2016). Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, 204(4):1353–1368.

Yang, Z. (1996). Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.*, 42:587–596.

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4):1811–1823.

Yang, Z. (2014). *Molecular Evolution A Statistical Approach*. Oxford University Press, Oxford, England.

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. USA*, 107:9264–9269.

Yang, Z. and Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12):3125–3135.

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci. U.S.A.*, 111(46):16448–16453.

Yu, Y., Ristic, N., and Nakhleh, L. (2013). Fast algorithms and heuristics for phylogenomics under ILS and hybridization. *BMC Bioinformatics*, 14 Suppl 15:S6.

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2018). Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.*, 35:504–517.

Chapter 3.4 The Concatenation Question

David Bryant
Department of Mathematics and Statistics, University of Otago
P.O. Box 56, Dunedin 9054 New Zealand
david.bryant@otago.ac.nz

Matthew W. Hahn
Department of Biology, Department of Computer Science, Indiana University
Bloomington IN 47405, USA
mwh@indiana.edu

Abstract

Gene tree discordance is now recognized as a major source of biological heterogeneity. How to deal with this heterogeneity is an unsolved problem, as the accurate inference of individual gene tree topologies is difficult. One solution has been to simply concatenate all of the data together, ignoring the underlying heterogeneity. Another approach infers gene tree topologies separately and combines the individual estimates in order to explicitly model this heterogeneity. Here we discuss the advantages and disadvantages of both approaches—using the gene trees singly or in concatenation—paying special attention to the sources of variance and implicit assumptions. We make it clear that all methods are likely to have their assumptions violated, though the consequence of these violations differs in different parts of parameter space. The main conclusion of our review is that different sources of error are more or less important in different settings, such that phylogenetics researchers should be using the methods most appropriate to their problems rather than stick to one dogmatically.

How to cite: David Bryant and Matthew W. Hahn (2020). The Concatenation Question. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 3.4, pp. 3.4:1–3.4:23. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Introduction

Phylogenetic inference is a hard problem, especially for deep divergences. There is the computational challenge: genomic data sets are huge and require sophisticated optimization and sampling algorithms. There is the statistical challenge: trees are awkward objects to do statistics on, and yet careful quantification of uncertainty is becoming more and more critical. There is the modelling challenge: access to full genome sequences has given us an appreciation of the complexity of evolutionary processes (see Chapter 2.1 [Simion et al. 2020]). Substitution rates vary across loci, across sites, across lineages and over time. Even the underlying Markov process can change (Inagaki et al., 2004; Jeffroy et al., 2006; Philippe et al., 2005, 2017).

The realization that different genes evolve under different substitution processes sparked the *total evidence* versus *consensus* debate in the 1990s (e.g. Page, 1996). The core point of disagreement was whether data should be analysed all together (total evidence) or separately and then combined (consensus; see Bull et al., 1993; De Queiroz, 1993). The debate spurred the development of new consensus methods, quartet methods, and tests for homogeneity. After a decade of sometimes bitter wrangling, the total evidence *versus* consensus debate died a natural death. The rise of efficient maximum likelihood and Bayesian software (see

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 3.4; pp. 3.4:1–3.4:23

A book completely handled by researchers.
No publisher has been paid.

Chapters 1.2 and 1.4 [Stamatakis and Kozlov 2020; Lartillot 2020]), and improved computers, made combined analysis with heterogeneous models tractable, placating both camps.

Heterogeneity in the substitution process is one thing; heterogeneity in tree topology is another. The fact that different loci can have different underlying histories has been recognised for a long time (Hudson, 1983; Tajima, 1983; Pamilo and Nei, 1988). However, it was not until coalescent theory leaked from population genetics into systematics (Maddison, 1997) that many realised how ubiquitous gene tree discordance could be. We can now be confident that much of the topological variation in recent divergences we see among loci is due to biological causes rather than technical errors (e.g. Brawand et al., 2014; *Heliconius* Genome Consortium, 2012; Fontaine et al., 2015; Novikova et al., 2016; Pease et al., 2016; Pollard et al., 2006; Rogers et al., 2019).

The heightened awareness of gene tree discordance appears to have revived the total evidence *versus* consensus debate, albeit in a different guise. Those on the (new) consensus side, armed with the multispecies coalescent, argue that gene trees should be inferred separately and then combined to infer a species tree (e.g. Edwards, 2009). Those on the (new) total evidence side pick holes in these arguments and remind readers of older arguments in support of concatenation (e.g. Gatesy and Springer, 2014).

If history repeats itself again, the revived debate will die a natural death. New, improved model-based methodologies will eventually address the concerns of both camps. In the interim, the debate motivates the discussion of fundamental issues related to the scale and scope of gene tree discordance, the importance of statistical consistency, and relative sources of error in phylogenetic inference. It also provides a convenient framework for a chapter discussing those issues.

In Section 2 we introduce much of the relevant theory about the population processes that lead to discordance, discussing incomplete lineage sorting, expected branch lengths, the "anomaly zone", and the impact of recombination. In Section 3 we discuss *summary tree methods*, approaches that estimate a gene tree for each locus and then combine these estimates. We discuss the consistency, and inconsistency, of these methods, and suggest that chopping the genome up into small chunks for separate analyses might open up these methods to systematic error and biases.

In Section 4 we examine the approach of concatenating all genes together before analysis, effectively ignoring the potential discordance among trees. This wholesale concatenation has been proven to be statistically inconsistent (Roch and Steel, 2015), though, as we point out, the branch lengths used in the proof are ridiculously short. The question of whether or not to concatenate is a question of finding a compromise between different kinds of error. The error from discordance exists for both recent and deep divergences, though with deep divergences it appears that alternative sources of error become much more important.

Section 5 examines one of the most confusing threads in the debate: whether or not we should tolerate recombination within loci. Oddly, both sides accuse the other of ignoring recombination and discordance. We discuss the arguments for and against combining trees, and some of the ways that have been proposed to overcome the bias associated with short loci. We also consider the pitfalls when considering clustered or binned genes as single loci in the multispecies coalescent.

In the final section we examine alternatives to summary methods and concatenated maximum likelihood, while also noting that there are excellent reasons for estimating individual gene trees, independent of species tree inference.

2 Gene tree discordance and the multispecies coalescent

One of the most important findings from genome-scale data in phylogenomics is that gene tree discordance is ubiquitous. Recognizing discordance between gene trees—and accounting for it in the inference of species trees—has been a major focus of the last decade of phylogenetic methods development. Among all of the possible causes of discordance, incomplete lineage sorting (ILS) has received the most attention, though introgression between species may well have a comparable real impact (Mallet et al., 2016). Here we focus on ILS, a concept we introduce in Section 2.2. While gene duplication and subsequent loss is also often included as a biological cause of discordance (e.g. Degnan and Rosenberg, 2009; Maddison, 1997), it is due to the mis-assignment of paralogs as orthologs (see Chapter 2.4 [Fernández et al. 2020]), and we do not consider it further.

2.1 Basic coalescent thinking

Two randomly chosen sequences at a locus from a single population share a common ancestor in the recent past. Under the Wright-Fisher model of diploid, hermaphroditic organisms with effective population size N, the probability that two autosomal sequences sampled from a single generation find a common ancestor in the previous generation is $1/2N$. We use $2N$ here because each of the N individuals carries two copies of this locus; alternatively, we can imagine a population of haploid individuals of size $2N$. A simple outcome of Mendelian inheritance is that the distribution of times back until two lineages find a common ancestor—that is, until they "coalesce"—is exponentially distributed with a mean of $2N$ generations. This process has a large variance, and independent loci sampled from the same two individuals will coalesce at many different times in the past.

Results for samples of size $n > 2$ can be derived under the n-coalescent model (Hudson, 1983; Kingman, 1982; Tajima, 1983). With $n = 3$ there are three equally probable topologies relating three lineages within a single population, and with $n = 4$ there are 18 equally probable labeled histories (by "labeled history" we mean that we distinguish between trees with the same relationships but that have lineages coalescing in a different temporal order). For all such topologies the coalescent model provides expectations for the times to coalescence, which in turn also imply branch lengths upon which mutations can accumulate (see Hein et al., 2004 and Wakeley, 2009 for an overview).

Importantly, in the n-coalescent model we assume that all observed mutations are neutral. This assumption allows us to completely separate the genealogical process of coalescence from the process by which mutations occur in the sample history. Every locus in this model has an underlying gene tree, irrespective of whether we are able to determine what it is from the pattern of informative mutations—our ability to infer a tree is not a necessary condition for its existence. More complex coalescent models than those described here are available, some incorporating selection, and some with non-Wright-Fisher populations (e.g. Spence et al., 2016). The importance of such models for phylogenetics is a largely unexplored area.

2.2 Incomplete lineage sorting

The small but finite amount of time it takes lineages to coalesce has significant consequences for variation in gene tree topologies. One useful way to think about this phenomenon is to ask whether all of the sampled lineages in a population have found their common ancestor before some pre-specified time in the past. Avise et al. (1983) referred to the case where all lineages find their common ancestor as "lineage sorting". Conversely, we now refer to

the case in which there are two or more lineages remaining as "incomplete lineage sorting" (ILS).

To be concrete, consider time measured in coalescent units, so that $T = t/2N$, where t is the number of generations. Given exponentially distributed coalescence times, the probability of lineage sorting of two lineages by time T in the past (i.e. the probability of 2 lineages going to 1 lineage) is:

$$P_{21}(T) = 1 - e^{-T}. \quad (1)$$

Likewise, the probability of incomplete lineage sorting (i.e. the probability of 2 lineages staying as 2 lineages) is:

$$P_{22}(T) = e^{-T}. \quad (2)$$

This result implies that only $\approx 63\%$ of loci will have coalesced by the mean expected time to coalescence ($2N$ generations, or 1 coalescent time unit), but that 95% of loci will have coalesced by $6N$ generations in the past. If we consider a species with an effective population size of 100,000 and a generation time of 1 year, these numbers imply that it would take on average 600,000 years for 95% of loci to sort. Similar calculations for the probability of lineage sorting among more than two lineages can also be made (Tavaré, 1984).

Incomplete lineage sorting is important for phylogenetics because it implies that, even when two species share an ancestral branch, not every gene tree sampled from those species will also have that branch. Enumerating the probabilities of specific topologies and their associated branch lengths in the presence of ILS is the goal of the *multispecies coalescent* model, which we discuss next.

2.3 The frequency of different topologies under the multispecies coalescent

Because of the time required for lineage sorting, ancestral populations that existed between closely spaced speciation events become very important. The multispecies coalescent (MSC) model (Hudson, 1983; Pamilo and Nei, 1988; Tajima, 1983; Takahata and Nei, 1985; Takahata, 1989) recognizes that coalescence in ancestral populations can determine the frequency and branch lengths of different topologies, and attempts to quantify these measures. The MSC is limited in many ways—it does not include many processes that can be modeled in the general coalescent framework—but it does provide an important guide to the effects of ILS on gene tree discordance.

Imagine that we have sampled one (haploid) individual from each of three species, A, B, and C, and that the true relationship between the species is $((A, B), C)$ (see Figure 1). We would like to know the probability of sampling a gene tree that matches this topology if we collect data from a single locus. Discordance at a locus can occur if the lineages sampled from A and B do not coalesce in their common ancestral population, and instead one of them coalesces with the lineage from C in the population ancestral to all three species. Regardless of how long the tip branches are (because no coalescence between species can occur along them), the probability that A and B do not coalesce in the most recent shared ancestral population is given by Equation 2, with T denoting the length of this internal branch. If there is no coalescence (i.e. if ILS occurs) then each of the three possible topologies are equally likely to occur in the common ancestral population of A, B, and C.

Under this model, the expected frequencies of the two discordant topologies are both (Hudson, 1983):

$$E[f_{((A,C),B)}] = E[f_{((B,C),A)}] = (1/3)e^{-T}. \quad (3)$$

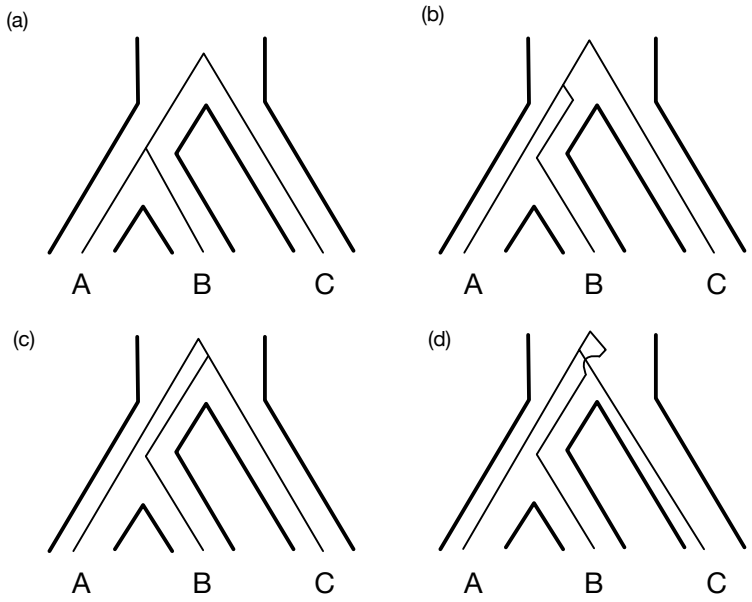


Figure 1 Incomplete lineage sorting and gene tree discordance. (a) Complete lineage sorting, so that the gene tree is consistent with the species tree. (b)-(d) Incomplete lineage sorting, of which only the first gives a gene tree consistent with the species tree.

A concordant topology is also produced under ILS, at the same frequency as the two discordant topologies (Hudson, 1983).

A concordant topology must be produced if there is lineage sorting (with probability $1 - e^{-T}$), so the total frequency of concordant topologies is:

$$E[f_{((A,B),C)}] = (1/3)e^{-T} + (1 - e^{-T}) = 1 - (2/3)e^{-T}. \quad (4)$$

We can see that there is more discordance with very small internal branch lengths (up to a maximum of 2/3 of all trees), and that at very long internal branch lengths there is essentially no discordance due to ILS. Following from the example given above, at $T = 6$ approximately 95% of loci will have sorted in the common ancestor of A and B, and will therefore be concordant. Of the remaining 5%, 1/3 will also be concordant, with the other loci equally split between the two discordant topologies.

Similar calculations can be made for arbitrarily large numbers of lineages undergoing ILS (Degnan and Salter, 2005). With four taxa undergoing ILS, there are now two internal branches of any species tree that must be considered, with ILS occurring in either one or both branches. While there are 18 possible labeled histories with four taxa, often only the 15 unlabeled histories are considered (e.g. Rosenberg, 2002), as we do not distinguish between, for instance, the two different possible sequences of coalescences in the topology $((A, B), (C, D))$ (either (A, B) first or (C, D) first). The number of possible topologies quickly explodes with more taxa. It is essential to realize, however, that these calculations reflect the number of lineages undergoing ILS, not the number of taxa in a tree. It may be that even in a tree of 100 taxa only 3 lineages are in a phylogenetic "knot" that induces ILS. In such cases we need only concern ourselves with ILS calculations for three taxa.

One of the most important take-home messages about the MSC is that ILS can occur at any time in the past. As can be seen from Equations 3 and 4, the only parameter determining the probability of discordance due to ILS is T, which measures the length of an

internal branch of the species tree in coalescent units (though note that most species trees are reported using absolute time or numbers of mutations per site per branch). Whether this internal branch existed 1 million or 100 million years ago, the amount of discordance due to ILS will be the same. However, our ability to determine a gene tree topology, and to ascribe it to ILS or not, *is* certainly dependent on how long ago these events occurred and how long the internal branches of individual gene trees are.

2.4 Gene tree branch lengths under the MSC

The expected branch lengths in both concordant and discordant gene trees are easily obtainable from the MSC model. As the coalescent process can have no effect on tip branch lengths after the most recent speciation event when one sequence is sampled from each species, total branch lengths will always have tip branch lengths added as a constant. Therefore, we focus on the expected lengths of gene tree branches above the tips.

When ILS occurs among three lineages, two coalescent events occur in the common ancestral population. Looking backwards, the first coalescence is expected to occur $2N/3$ generations in the past (Figure 1(b-d)). Following this event, which is equally likely to join any of the three possible pairs of sequences, there are two lineages left in the tree, and therefore an average of $2N$ more generations until the entire sample has reached its most recent common ancestor. This result means that the internal branch of any topology that is the result of ILS (whether concordant or discordant) has an expected length of $2N$ generations. Even in "hard polytomies", where the length of the internal branch of the species tree is zero, each of the three gene trees has an internal branch with expected length $2N$. In contrast, the internal branch of loci that complete the lineage sorting process have a minimum expected length of $2N$ generations and a maximum length equal to the length of the internal branch of the species tree. Mendes and Hahn (2018) provide analytical formulas for these expectations.

To give some perspective on how long such an internal branch is, recall that the expected number of pairwise differences between two sequences within a population is $4N\mu$, where μ is the per-nucleotide mutation rate (often this compound parameter is referred to as θ). The expected number of mutations on the internal branch of discordant trees is $2N\mu$, or about half the number of pairwise differences. As the proportion of sites with such differences are generally at or below 1% for multicellular organisms (Leffler et al. 2012), we can begin to understand why it is so hard to accurately identify discordant gene trees. Note that this internal branch of discordant trees has the same short length no matter how far back in time the ILS occurred and no matter how long the internal branch of the species tree is. This is why discordant gene trees due to ILS will always have lower bootstrap support than any concordant tree—because they always have a shorter internal branch—and why using a bootstrap cut-off to determine which gene trees to include in an analysis could result in a biased estimate of discordance.

2.5 The anomalous anomaly zone

Under the multispecies coalescent, each species tree determines a distribution on the set of gene trees. There have been several theoretical results on this distribution, mainly for the case where exactly one individual is sampled from each species. It is tempting to think of this distribution as a cloud of random gene trees, centred on the species tree. Degnan and Rosenberg (2006) showed that this picture can be misleading. One can construct a species tree such that the most likely gene tree under the multispecies coalescent is *not* the species

tree. This observation, perhaps more than any other, has been used to justify methods that account for incomplete lineage sorting in inferring the species tree.

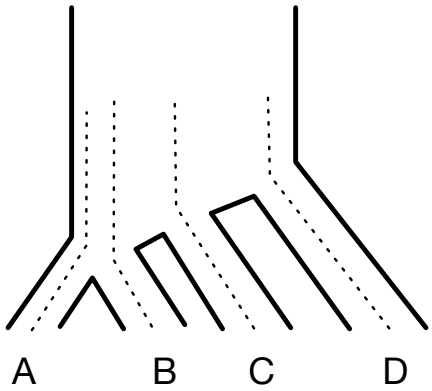


Figure 2 Tree used to generate an anomaly zone in Degnan and Rosenberg (2006). The branches below the root are short enough that almost all coalescent events occur in the population at the root. As the coalescent process in one population results in more trees that are balanced (e.g. $((A, B), (C, D))$), these gene trees will have higher probability under the MSC than the underlying species tree.

Consider the species tree in Figure 2. If the two successive internal branches in this species tree are short enough (in coalescent units), coalescences are most likely to occur in the ancestral population of all four lineages. Under the coalescent model in a single population, symmetric trees, like ((A,B),(C,D)) have higher probability than asymmetric ("caterpillar") trees like (((A,B),C),D) because the former are associated with two labeled histories while the latter are associated with only a single labeled history. As a result, if the species tree is a caterpillar and most coalescence occurs in the single ancestral population, an asymmetric gene tree matching the species tree can be less common than one of the symmetric gene trees. The term "anomaly zone" is used to describe the area of parameter space (in terms of branch lengths in the species tree), where the gene tree concordant with the species tree is less probable than some other tree.

Despite the anomaly zone's bogeyman-like status in phylogenetics, it is not as scary as it looks. For instance, there is no anomaly zone for gene trees with branch lengths. The times between coalescent events have an exponential density. The density function of an exponential random variable is strictly decreasing, so gets larger for values close to zero. The mode, or most "likely" value, for an exponential random variable is zero, and the most likely time of coalescence is zero or effectively instantaneous. As a consequence, the mode of the distribution of gene trees with branch lengths under the MSC is identical to the species tree! For the same reason, in the multispecies coalescent with extremely small population sizes (such that pairs of lineages coalesce as soon as possible), every gene tree would match the species tree exactly.

Nevertheless the anomaly zone *does* cause problems with methods for inferring species trees based on counts or frequencies of gene trees. It was also thought that the anomaly zone was responsible for consistency problems with concatenated maximum likelihood, though as we will see later it is not the anomaly zone that is responsible.

2.6 The coalescent with recombination

The classical MSC model makes two important assumptions about the role of recombination, neither of which is likely to be true in real data but both of which are required to produce the topological distributions expected under the MSC. The first assumption is that we are dealing with individually non-recombining loci, such that each locus or gene contains only a single underlying topology. Non-recombining loci such as mtDNA, cpDNA, or the Y (or W) chromosome all conform to this assumption. For sequences drawn from the autosomal nuclear genome, the length of non-recombining loci is a function of rates of recombination and population sizes (N).

These considerations raise the question of how long we expect non-recombining loci in the nuclear genome to be. The rate of recombination varies along the genome and across species. In humans, the average length of non-recombining autosomal loci is 4.8-5.9 kilobases (International HapMap Consortium, 2005), though there is a huge variance in the length of such blocks. For species with larger population sizes such as *Drosophila*, there is more effective recombination and block sizes are commensurately smaller, on the order of hundreds of bases or less (Hey and Nielsen, 2004). However, because the amount of nucleotide diversity also scales with population size, the large block sizes in species with small populations like humans do not necessarily result in more phylogenetic resolution within each locus.

While a strict interpretation of the MSC assumes non-recombining loci, there are some kinds of recombinations that have no effect on inference. If the recombination is limited to lineages within a branch, different sites will have identical gene trees, even if they are undergoing recombination.

The real concern of intra-locus recombination for inference is when there are multiple histories present among loci, as when there is incomplete lineage sorting or introgression. When this occurs, different sites within a single protein-coding gene can have discordant gene trees. In fact, Mendes et al. (2019) showed that 70% and 91% of protein-coding genes in primates and *Drosophila*, respectively, contain two or more gene tree topologies from a single phylogenetic knot (i.e. three species undergoing ILS). Even if genes were on average the same length as non-recombining stretches of chromosome, multiple trees will be combined unless the recombination events exactly flank the sequence being used for inference. When there are multiple knots across a larger tree, the length of loci that do not have a single recombination event within them at any point in the tree can become vanishingly small (Gatesy and Springer, 2014). We return to the issue of intra-locus recombination below, as it affects all of the methods we discuss here.

There is a second critical assumption that the MSC makes about recombination: that different loci have independent gene trees (conditional on the species tree). In other words, it assumes that there is sufficient recombination between loci that gene trees for different loci are independent (conditional on the species tree). Non-independence of samples is a common problem across statistics, known to cause greater variability than expected (overdispersion). The consequences of assuming independence in phylogenetics are not well understood, but generally assumptions of this type result in greater confidence in results than is warranted.

3 Summary gene tree methods

In a summary tree method, separate gene trees are estimated for each locus, and these gene trees are then used to infer the species tree. In the contemporary revival of the total evidence *versus* consensus debate, those advocating summary tree methods fall squarely in the "consensus" camp. Examples of this approach include ASTRAL (Mirarab et al.,

2014), MP-EST (Liu et al., 2010), and ASTRID (Vachaspati and Warnow, 2015), with new methods and updates appearing monthly. Summary tree methods are sometimes referred to as "shortcut" coalescent methods, as the full likelihood of the species tree under the MSC is bypassed in favor of simple expedients.

3.1 Non-anomalous subtrees

Most existing summary tree methods depend heavily on a couple of key results regarding rooted triples and unrooted quartets of gene trees. In the anomaly zone the most probable gene tree under the MSC can be different than the species tree (ignoring branch lengths). Oddly enough, if we throw away taxa, the anomaly zone disappears. To demonstrate this phenomenon, consider the two following properties of the MSC:

1. In a rooted species tree with three taxa and one individual sampled per species, the most probable rooted gene tree is the same as the species tree, regardless of internal branch length.
2. In an unrooted species tree with four taxa and one individual sampled per species, the most probable unrooted gene tree is the same as the species tree, regardless of internal branch length.

Both of these observations are direct consequences of calculations from Tajima, Hudson, and Nei dating back 30-40 years; see Degnan and Rosenberg (2006) and Degnan (2013) for recent derivations.

At first glance, these properties of the MSC appear to create a paradox. If we consider all taxa at the same time, the most probable gene tree need not be the same as the species tree. However, if we consider every triple (in the rooted case) or quartet (in the unrooted case) then the most probable small trees will match the species tree, even though any rooted tree is determined by its triples and every unrooted tree is determined by its quartets. The explanation for this is that even the wrong trees might have some of the correct triples or quartets, and the probability of observing a particular triple combines the probability of observing it when the gene tree equals the species tree and the probability of observing it when it does not.

The absence of non-anomalous gene trees with four taxa (or three in the rooted case) makes it easy to design species tree methods that are consistent under the MSC. Given a sample of unrooted gene trees, we determine the most frequent four-taxon trees for each subset of four taxa. As the number of loci increases, the probability of inferring the four-taxon tree concordant with the speces tree approaches one. Reconstructing the species tree from its quartets is straightforward. The case for rooted trees is the same, only there we deal with rooted triples (three-taxon trees) rather than quartets. The term "coalescent aware" was coined for methods which made use of these results, even though they do not necessarily require any coalescent calculations.

The realisation that methods using subtrees of larger trees are apparently immune to gene tree discordance led to a real 90s-style revival in phylogenetic methodology. A general approach that had, for the most part, fallen into disuse, suddenly became a mainstream tool in systematics and even required by some journal editors.

3.2 Inconsistency of summary tree methods

The concept of *statistical constistency* features prominently in the discussion and promotion of summary tree methods. In general, an estimator for a quantity is statistically consistent if the probability of it returning the correct value converges to one as the amount of data

increases, given that the data are generated by the assumed model. Thus a phylogenetic method is consistent if the method converges to the correct tree when there is no model violation and the amount of data (number of sites or loci) goes to infinity.

Current summary tree methods make the fundamental assumption that the inferred gene trees have the same distribution as do gene trees under the MSC. If the distribution is the same, methods based on rooted triples and unrooted quartets will infer the underlying species tree accurately, given sufficiently many loci. However in practice we do not have access to the gene trees themselves, but only estimates of the gene trees. Those estimates, particularly in deep phylogenetics, can be error prone.

For these reasons, Warnow (2015) describes two flavours of statistical consistency for the multispecies coalescent. The *weak* version corresponds to consistency conditional on correctly inferred gene trees. The *strong* version says that the estimated species tree will converge in probability as the number of loci increases even with a bound on the length of each locus. In practice, it is strong consistency that is relevant to phylogenetics.

Summary methods do not typically satisfy strong consistency. This is not that surprising—it would be a minor miracle if the distribution implied by the multispecies coalescent happened to line up exactly with the distribution resulting from inferring trees with sampling error. Roch et al. (2018) provide a formal proof of inconsistency. Here we will settle for an informal argument, one which illustrates a particularly important point. A similar phenomenon is documented by Wang et al. (2019).

Maximum likelihood (ML) is biased on finite sequences; it is, after all, a non-linear estimator. The most extensively studied example of bias in ML is "long branch attraction", which describes the zone of parameter space in which long, non-sister branches are erroneously inferred to be sister (e.g. species A and D in Figure 3). This is widely known as the Felsenstein zone (Felsenstein, 1978). Though it is usually thought of as a larger stumbling block for parsimony, when the mix of branch lengths is sufficiently extreme, and sequences sufficiently short, the Felsenstein zone is also challenging for ML.

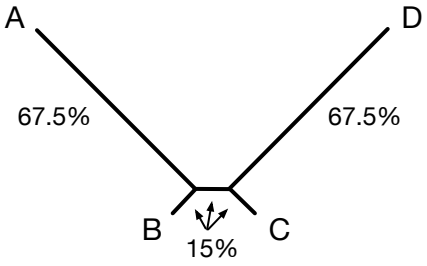


Figure 3 Four taxon tree on which ML is biased with short sequences. Branch lengths are in expected percentage of non-identical sites (Adapted from Swofford et al., 2001).

Swofford et al. (2001) studied the problem of inferring trees simulated on the tree in Figure 3. The long branches have about 1.7 expected substitutions per site while the short branches have about 0.16 expected substitutions per site. If you simulate sequences of length 50 base pairs on this tree, then the maximum likelihood tree will be $((A, D), (B, C))$ with probability 41%, $((A, B), (C, D))$ with probability 34% and $((A, C), (B, D))$ with probability 25%. This implies that the maximum likelihood tree is correct only 34% of the time.

Now suppose that the tree in Figure 3 is the species tree, and that population sizes are so small (or branch lengths so long) that there is negligible ILS. If the loci only have 50 sites each, then, as the number of loci increases, the frequency of the correct tree will converge

on 34%, while the frequency of the incorrect tree will converge on 41%. Any of the standard summary tree methods will then select the wrong tree as the species tree with certainty as the number of loci increases. In other words, *summary tree methods are statistically inconsistent, in the "strong" sense.*

A method can be statistically inconsistent and yet perform well in practice. In itself, a proof of consistency or inconsistency only tells us a limited amount because it only addresses asymptotic bias. In any statistical estimation problem, phylogenetic inference included, there is a trade-off between bias and variance. An estimator might have some bias associated with it, even with infinite data; however, it can be advantageous to just live with that bias if the overall level of error can be kept under control. The real value of proofs of inconsistency, or indeed of simulations demonstrating bias, is that they help us to identify contexts within which a method might be misleading. The classic proof of Felsenstein (1978) that parsimony is inconsistent is useful because it identifies important and real situations where the method can be misleading. It also helped to identify similar problem areas for other methods, including ML (e.g. Kim 1996).

3.3 The problem with summary tree methods

Summary tree methods can be inconsistent because maximum likelihood is biased. In the simple example we gave above (Figure 3), maximum likelihood would select an incorrect gene tree more often than the correct tree. The potential for long branch attraction, and other forms of bias, increases as phylogenies get deeper and the variation in evolutionary processes gets more complex. Variation in substitution rates among sites and across the tree make it difficult to correct for homoplasy and multiple substitutions.

Extensive work has been done examining these issues, and how they can affect phylogenetic inference (Philippe and Roure 2011; Philippe et al. 2011, 2017; Chapter 2.1 [Simion et al. 2020]). Not only can long branch attraction be difficult to diagnose, it can lead to a complete reshuffling of the inferred tree. This stands in contrast to ILS, which typically only has a local impact on the topology around short internal branches.

The standard strategy for dealing with heterogeneity in the substitution process is to try to construct generative models of how the processes can change. There are many advantages to using models, particularly the fact that we can start to understand features of the actual substitution process and their impact on inference.

Obviously, then, we would like to apply model-based approaches to the inference of gene trees. However, in order to fit complex models and to carry out reliable inference using these models, we need long sequences. We need longer sequences because modelling variation will inevitably result in increased sampling variance and small-sample bias. In the example of Figure 3 it took slightly more than 50 sites to overcome the bias. For larger, deeper trees, and multi-faceted, complex models, it could take many many more sites. Kück et al. (2012) report alignments where maximum likelihood is still biased with over 100,000 sites!

By chopping up the genome and analysing each fragment independently, summary tree methods run the risk of substantial and systematic bias, replicated independently for each locus. The obvious solution to this problem is to join loci together to make longer alignments, but this approach has pitfalls of its own. It is these problems that we discuss next.

4 Concatenation

Standing in the opposite corner from summary tree analysis stands the total evidence, or concatenation, approach. In this approach all loci are analysed together, using models that

incorporate complex substitution processes. We call this approach *concatenated ML*. The advantages of concatenated ML are that all of the models and technologies developed for deep phylogenetics can be applied to the concatenated alignment. The sequence lengths are long enough to fit the models with some semblance of reliability, which ensures more accurate inferences for large areas of parameter space.

4.1 Inconsistency of concatenated ML

The problem with concatenating all of the loci is that concatenated ML is not statistically consistent on data generated by the MSC, a fact that has been used repeatedly to justify summary gene tree methods. In general, if a parameter is identifiable, and you use a maximum likelihood estimator with the correct model for the data, then a maximum likelihood estimator is consistent. As a consequence, concatenated ML is consistent on data generated on a single gene tree. However, it is not consistent on data generated on discordant gene trees (or a mix of concordant and discordant trees). Given the existence of incomplete lineage sorting and the anomaly zone, it should perhaps come as no surprise that maximum likelihood has trouble on alignments containing many conflicting gene trees.

The existence of the anomaly zone does not, by itself, imply statistical inconsistency of concatenated ML under the MSC. Nevertheless, simulation studies (Kubatko and Degnan, 2007) indicate that the short internal branches of caterpillar trees associated with the anomaly zone cause problems for maximum likelihood estimation on concatenated data. Mendes and Hahn (2018) have shown that concatenated ML can fail both inside and outside the anomaly zone, as the reasons for the inconsistency are not directly driven by the identity of the most probable gene tree. Indeed, both parsimony and neighbor-joining appear to perform suprisingly well in the presence of large amounts of discordance, on species trees both inside and outside the four-taxon anomaly zone (Liu and Edwards, 2009; Mendes and Hahn, 2018).

The first analytical proof of the inconsistency of concatenated ML under the MSC was provided by Roch and Steel (2015). A nice overview of the proof can be found in Warnow (2015). As with the anomaly zone proof, the basic issue is that when internal branch lengths are very short almost all of the coalescent events occur in the ancestral population. Because gene trees generated under the coalescent model for a single population favour one kind of tree over others, the gene trees resulting from species trees with very short internal nodes also favor one (wrong) topology over the others (Figure 4).

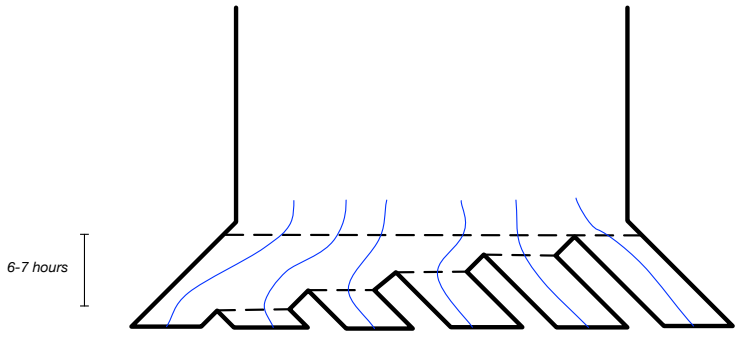


Figure 4 The species tree for which concatenated ML fails in Roch and Steel (2015). The times between speciation events are rather short: for effective population size $N = 100,000$, $\theta = 0.001$ and a generation time of one year, consecutive speciation events are separated by a little under 100 minutes.

To illustrate how short the internal branch lengths have to be for concatenated ML to fail in Roch and Steel's proof, it is useful to consider the parameter values required by their construction. Suppose that there are n taxa, and let β be an upper bound on the length of any internal branches, measured in coalescent units. Claims 7 and 5 in Roch and Steel (2015) then imply that β satisfies

$$\left(e^{-\binom{n}{2}\beta}\right)^n e^{-\theta(n^2\beta)} \geq 1 - \frac{\theta^2}{n}.$$

If we plug in $n = 6$, which is the number of taxa used in their construction, we have $\beta < 1.86 \times 10^{-9}$ for $\theta = 0.001$. To put this figure in context, suppose that we are considering a species with an effective population size of around $100,000$ and generation time of 1 year. The number of generations along each internal branch is then bounded above by $1.86 \times 10^{-9} \times 100,000$, corresponding to 1.86×10^{-4} years, or just under 100 minutes. In other words, Roch and Steel have proven that concatenated ML is inconsistent on a species tree where the gap between divergences is less than 100 minutes, or 6-7 hours total for all five speciation events to occur.

We are being a little facetious here. The inconsistency result is correct and, moreover, we believe that concatenated ML will still continue to be inconsistent on more reasonable species trees (if difficult to prove analytically). However, there is also a serious side. Just because a method is inconsistent for some parameter values does not mean that it is not the best methods for others. We contend that this may well be the case for concatenated ML: when the amount of incomplete lineage sorting is small, perhaps relative to other sources of noise, concatenated ML performs extremely well (Mirarab et al., 2016). When the amount of ILS is large, then we should be concerned about the shortcomings of concatenated ML.

4.2 Balancing different sources of error

When we carry out concatenated ML analysis in the context of the MSC we are committing the sin of model misspecification: the model used for inference does not match the one that generated the data. Model misspecification is, of course, widespread in statistics—indeed there are few statistical analyses where model assumptions all hold exactly. The key issue is the extent to which model misspecification is misleading in practice and in context. Will the model misspecification completely rearrange the tree, or just locally distort the topology around a few tiny branches?

Context is particularly important. In phylogenetics different sources of error play quite different roles at different depths in the tree. Simulations, and theory, demonstrate that phylogenetic inference via concatenated ML is badly misled when branch lengths are very short, resulting in high levels of ILS. The error from ignoring ILS leads to local rearrangements in the tree, rather than global errors. Furthermore, in the area of parameter space where such errors occur, $< 15\%$ of all gene tree topologies match either the "true" species tree or the tree returned by concatenated ML (Kubatko and Degnan, 2007; Degnan and Rosenberg, 2006). In other words, no matter which tree is chosen, $\sim 85\%$ of loci in the genome will have a different evolutionary history. If rearrangements are local and all answers represent only a small minority of gene trees, then this source of error is far more benign than what we may expect from substitution rate errors and long branch attraction.

We can get a sense for the relative contribution of different sources of error by comparing the corresponding sources of variance when estimating genetic distances between species. Consider sequences sampled from two species separated t generations ago. Under a simple Poisson mutation model, the number of nucleotide differences between two sequences has

expectation

$$2t\mu + \theta \tag{5}$$

and a variance of

$$2t\mu + \theta(1 + \theta) \tag{6}$$

(Gillespie and Langley, 1979). Here $2t\mu$ is expected number of mutations since the time of species divergence, and $\theta = 4N\mu$ is again the expected number of differences between two sequences sampled from the ancestral population at the time the lineages split.

The two terms in Equation 6 correspond to variance from the mutation process since divergence and variance from the coalescent process prior to divergence. How do these quantities compare in practice? A typical value for θ, and hence for the coalescent variance ($= \theta(1+\theta)$), might be around $\theta = 0.01$. The value for mutational variance scales with t, and therefore depends on how long ago the species split. If $t \approx 2N$ then the two contibutions to variance are almost identical: the variability from the coalescent process matches the variability from the mutational process. Hence the coalescent will be a major source of error. In a species with a generation time of one year and $N = 100,000$ this corresponds to a divergence time of $200,000$ years.

As t increases the variance from mutation also increases, but the contribution from the coalescent stays the same. So at 1 million years, the variance of mutation will be five times that of the coalescent; at 10 million years it will be 50 times. Deeper than 20 million years, and the coalescent will contribute less than 1% of the variance. In effect, the misspecification resulting from ignoring the coalescent should have little or no impact on inference (at least for divergence times).

There are some obvious limitations in this example. For one thing, relative variances would change with more sequences, or more complex models. The key fact, though, is that mutational variance increases significantly with depth of divergence, whereas coalescence variance is the same at any depth. We expect coalescence to play a more significant role when unravelling recent divergences, but to be swamped by other sources of error when examining deeper divergences.

Two recent studies provide empirical evidence that mutational variance and the modelling error that comes along with it can dominate the inference of gene tree discordance in deep phylogenies. Richards et al. (2018) carried out a detailed and careful phylogenetic analysis of genes in the mitochondrial genome for several deep clades of vertebrates. Recombination in the mitochondria is rare, at least relative to autosomal recombination (White et al., 2013). Hence, any discordance observed among inferred trees is most likely to be a consequence of phylogenetic error rather than biological gene tree heterogeneity. Nevertheless, the study observed gene tree conflict at a level commensurate with that observed in nuclear genomes. While there was no "consistent" discordance, as there would be under ILS, the observation of strongly supported discordant trees is worrying.

Scornavacca and Galtier (2017) employed results on the expected length of internal branches of discordant trees to put an upper bound on the expected proportion of sites affected by ILS. Using a rough estimate of θ from extant mammals, they show that the observed proportion of sites supporting a gene tree discordant with the species tree is far higher than that expected. This result was stronger for deeper nodes in the placental mammal phylogeny, suggesting that only a small fraction of discordant sites can be explained by ILS deeper in the tree. In this context, other sources of error are swamping ILS.

In summary, then, there is no question that ILS is an important cause of gene tree discordance, especially when looking at recently diverged populations or species. It is not clear, however, that the phylogenetic "error" due to ILS trumps all other sources of error, especially as we move into the more distant past. It is not that problems due to ILS get smaller, only that the error due to all other causes gets much larger.

5 The role of recombination in the debate

5.1 Concatalescence

One issue that we have mostly avoided discussing so far is whether the loci analyzed by summary tree methods are themselves non-recombining. A standard unit of phylogenetic analysis is the protein-coding gene. As mentioned earlier, the vast majority of protein-coding genes in eukaryotic genomes are likely to contain two or more topologies in the presence of ILS. While single exons less often contain multiple topologies (Mendes et al., 2019), they are much shorter and are therefore both less likely to be able to fully resolve trees containing many taxa and will provide many fewer sites with which to fit complex substitution models. The implicit compromise of using even single protein-coding genes is that we have enough sequence with which to carry out "good enough" phylogenetic analyses, even though we may be violating the MSC. This compromise approach has been given the portmanteau "concatalescence" (Gatesy and Springer, 2014).

There has been quite a kerfuffle in the literature surrounding concatalescence (Gatesy and Springer, 2014; Liu et al., 2014b; Springer and Gatesy, 2016; Edwards et al., 2016). It seems to us that the main question is not whether current approaches using single protein-coding genes violate the MSC—they almost certainly do—but what effect this violation has on the inferred topologies, and especially the distribution of inferred topologies.

We have already seen that in extreme cases of ILS, concatenated ML will converge on the wrong tree with more and more data (Kubatko and Degnan, 2007). What is less clear is the behavior of shorter genes that combine only a handful of different topologies. Some simulations have been done to examine the effect of recombination on realistic gene lengths (Lanier and Knowles, 2012), finding little effect relative to other sources of phylogenetic error. However, these simulations have been criticized as having too little recombination (because the intervening introns were not taken into account; Gatesy and Springer, 2014), and did not seem to include the areas of parameter space where concatenated ML fails. If individual gene trees are biased by concatenation, then so too will be the rooted triplets and unrooted quartets extracted from them for use with summary methods.

Regardless of the criticisms of published simulations, researchers in favor of summary gene tree methods face an apparent paradox: if typical protein-coding genes are immune to the effects of recombination, ILS, and concatenation (and can therefore be used to construct gene trees), then why not concatenate all the loci? Unfortunately, no theory yet exists that bounds the amount of recombination and ILS allowable while still producing correct trees. Further work is clearly needed to know how far such methods can be pushed.

5.2 Conditional concatenation and binning

Summary methods are problematic because ML is biased. We have now seen two causes for this bias: sequences that are too short to accurately model the substitution process (section 3.3) and sequences that are so long that they contain multiple conflicting topologies within them (section 5.1). In the next section we discuss "full" likelihood methods that can possibly

deal with the latter problem, and here we address strategies that have been used to increase the length of loci used as input to summary methods.

One strategy is that of conditional concatenation, in which loci are combined for analysis only if they pass some kind of test of concordance. Conditional concatenation has a long and well-cited history, and featured in the total-evidence *versus* consensus debate (e.g. Bull et al., 1993; De Queiroz, 1993). There are many different topology-based congruence tests available (see Leigh et al. 2011 for a comprehensive review), but many pitfalls to these approaches as well. As we mentioned earlier, there are potential biases in the bootstrap support for gene trees in the MSC: the edges in gene trees that are discordant with the species tree are likely to be short, with length determined by within-population coalescence. Hence bootstrap support for discordant trees may well be systematically lower than for concordant trees. Ironically, discordant trees will then be more likely to be combined with other trees than concordant trees.

More seriously, there is a fundamental problem with using topology-based tests to determine whether alignments can be combined: the main reason we are considering concatenation in the first place is because we cannot reliably construct phylogenies for single genes. How then can we expect these inferred single-gene trees to reliably inform us about phylogenetic dependencies? It may well be that character-based tests of incongruence can sidestep this issue, or at least appear to. However, since the objective is to emulate a non-recombining locus, it may well be more appropriate to apply one of the dozens of tests for recombination instead (e.g. Martin et al., 2010).

There has also been a lot of confusion in the literature about the interpretation of these combined genes. What is clear is that the combined genes should in no way be considered to be a linked, non-recombining locus with respect to the MSC. Indeed, there is no guarantee that the combined genes are even on the same chromosome. One possible interpretation of these conditionally concatenated loci is that they evolved separately along the species tree, but happen to come from gene trees that are not significantly different. After all, different gene trees are independent only conditioned on the species tree, and it is therefore no surprise that unlinked loci might have similar gene trees. By concatenating the genes, we are taking advantage of the fact that the gene trees are from the same species tree and so are interdependent, allowing us to fit more sophisticated and robust models. Once that inference process has completed, the genes should be considered independent with respect to the MSC—they just happened to have their gene trees estimated at the same time. This separation is implicit in "weighted statistical binning" (Bayzid et al., 2015). The combination of a sophisticated phylogenetic concordance test with a strategy like weighted statistical binning may offer a compromise choice that hits the "goldilocks" zone for multi-gene inference, though there is still plenty of work to do in understanding the systematic biases this could introduce.

6 Beyond summary methods *versus* concatenation

6.1 Full phylogenetic methods incorporating ILS

Summary gene tree methods are not the only way to incorporate gene tree heterogeneity into phylogenetic inference, and almost certainly are not the best way (though they are fast). Several methods exist that can carry out exact likelihood (or similar) calculations under the MSC, using these calculations to infer species trees from data (see Chapter 3.3 [Rannala et al. 2020]). Although these methods overcome many of the problems associated with summary approaches, they still face some of the same issues, especially those associated

with recombination.

The most widely used methods can be conveniently separated into two groups: those that use blocks of sequence, but assume no recombination within loci and free recombination between loci, and those that use only variable sites, but assume that there is free recombination between them. In the first category are methods implemented in BPP (Rannala and Yang, 2017), StarBeast (Heled and Drummond, 2009; Ogilvie et al., 2017) and PhyloNet (Wen et al., 2018). All three methods can work directly from individual gene alignments, calculating the likelihood of the data under the MSC. They accommodate sampling error in the gene trees that summary tree methods ignore. These methods (or their extensions; Zhang et al., 2018) are also able to infer species networks—essentially the species tree with reticulations—though the methods that do so require time-consuming MCMC sampling. Regardless of the way in which tree space is explored, these methods still assume that each input gene is a non-recombining unit, and therefore face some of the same modelling questions as summary methods.

In the second category are methods that assume free recombination between individually varying sites. Methods that use this type of data are varied, including SNAPP (Bryant et al., 2012), PoMo (De Maio et al., 2015), and SVDquartets (Chifman and Kubatko, 2014). While SNAPP and PoMo are optimal for species with recent splits (and multiple individuals sampled per species), SVDquartets is able to infer species trees with deep splits. These methods all avoid the issues with short non-recombining blocks of sequence, completely circumventing the problem by combining together a large number of independently evolving loci. Although complex substitution models incorporating all the different kinds of rate variation observed are not yet included in the tools listed here, these methods are some of the most promising for the future of phylogenetics.

6.2 Why genes should still be analysed separately

Even if you believe that species tree inference should only be carried out with concatenated data, it is still useful to infer trees for each gene. Arguments in favor of the examination of individual gene trees go back as far as the consensus/total-evidence debate (e.g. De Queiroz, 1993), and genomic data has only made this more true. Individual gene trees can reveal an enormous amount about variation in history along the genome, different rates of evolution in different genomic compartments, and different potential biases or patterns in a dataset. The signal in the data is in the variable history among loci, not just species relationships (Bravo et al., 2019).

One obvious example of where the study of individual gene trees can help is in cases of horizontal gene transfer (HGT) or introgression between species. The disagreement among trees is widely considered to be the best evidence for transfer (Soucy et al., 2015). Similarly, gene flow between sexually reproducing species can result in gene tree discordance at introgressed loci. The distribution of discordant trees along the genome is one of the few indications that introgression is occurring (e.g. Liu et al., 2014a), and the distinct heights and branch lengths of introgressed trees can help to disentangle complex histories (e.g. Fontaine et al., 2015; Kearns et al., 2018).

Because genes underlie traits, gene trees may also be a much better guide to trait evolution than species trees, especially when there is a lot of discordance (Hahn and Nakhleh, 2016). In cases with extreme levels of discordance, such as adaptive radiations, it may even be possible to associate individual discordant loci with incongruent traits (e.g. Pease et al., 2016; Wu et al., 2018). Radiations may be one of the best arguments for approaches that examine individual gene trees, as it becomes highly unlikely that *any* locus follows the inferred

species history (e.g. Jarvis et al., 2014).

Finally, the visualization of gene tree heterogeneity may itself be a worthwhile endeavor. Hillis et al. (2005) showed how a collection of inferred gene trees could be visualized in tree space using multidimensional scaling. Duchêne et al. (2017) used this multidimensional scaling approach to identify clusters of gene tree topologies supporting conflicting resolutions of the species tree, and were able to show that the clusters were generated by ILS. Other sorts of visualization tools may be equally useful in different contexts (e.g. Esser et al., 2004).

6.3 Moving forward

Phylogenetic inference is a hard problem, especially for deep divergences. As we have seen, much of the difficulty stems from how and what to model, and the extent to which different models impact on our inference.

Therefore, the choice of methods to use should be informed by the largest sources of error. At shallower timescales gene trees can be accurately inferred and ILS (and introgression) can be large sources of variance among gene trees. At deeper timescales the sources of variance flip, such that ILS becomes relatively less important. ILS certainly occurs at deep timescales, but many other processes also come into play, making the inference of individual gene trees much harder. While we hope that researchers interested in resolving relationships at, for instance, the base of animals keep the possibility of gene tree discordance in mind, it is certainly understandable that the methods they employ to infer a species tree do not model this process explicitly.

References

Avise, J. C., Shapira, J. F., Daniel, S. W., Aquadro, C. F., and Lansman, R. A. (1983). Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. *Molecular Biology and Evolution*, 1:38–56.

Bayzid, M. S., Mirarab, S., Boussau, B., and Warnow, T. (2015). Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLoS One*, 10(6):e0129183.

Bravo, G. A., Antonelli, A., Bacon, C. D., Bartoszek, K., Blom, M. P. K., Huynh, S., Jones, G., Knowles, L. L., Lamichhaney, S., Marcussen, T., Morlon, H., Nakhleh, L. K., Oxelman, B., Pfeil, B., Schliep, A., Wahlberg, N., Werneck, F. P., Wiedenhoeft, J., Willows-Munro, S., and Edwards, S. V. (2019). Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. *PeerJ*, 7:e6399.

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A. Y., Lim, Z. W., Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H. J., Russell, P., Aken, B., Alfoldi, J., Amemiya, C., Azzouzi, N., Baroiller, J.-F., Barloy-Hubler, F., Berlin, A., Bloomquist, R., Carleton, K. L., Conte, M. A., D'Cotta, H., Eshel, O., Gaffney, L., Galibert, F., Gante, H. F., Gnerre, S., Greuter, L., Guyon, R., Haddad, N. S., Haerty, W., Harris, R. M., Hofmann, H. A., Hourlier, T., Hulata, G., Jaffe, D. B., Lara, M., Lee, A. P., MacCallum, I., Mwaiko, S., Nikaido, M., Nishihara, H., Ozouf-Costaz, C., Penman, D. J., Przybylski, D., Rakotomanga, M., Renn, S. C. P., Ribeiro, F. J., Ron, M., Salzburger, W., Sanchez-Pulido, L., Santos, M. E., Searle, S., Sharpe, T., Swofford, R., Tan, F. J., Williams, L., Young, S., Yin, S., Okada, N., Kocher, T. D., Miska, E. A., Lander, E. S., Venkatesh, B., Fernald, R. D., Meyer, A., Ponting, C. P.,

Streelman, J. T., Lindblad-Toh, K., Seehausen, O., and Di Palma, F. (2014). The genomic substrate for adaptive radiation in african cichlid fish. *Nature*, 513(7518):375–381.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29(8):1917–1932.

Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L., and Waddell, P. J. (1993). Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42(3):384–397.

Chifman, J. and Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, 30(23):3317–3324.

De Maio, N., Schrempf, D., and Kosiol, C. (2015). PoMo: an allele frequency-based approach for species tree estimation. *Systematic Biology*, 64(6):1018–1031.

De Queiroz, A. (1993). For consensus (sometimes). *Systematic Biology*, 42(3):368–372.

Degnan, J. H. (2013). Anomalous unrooted gene trees. *Systematic Biology*, 62(4):574–590.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5):e68.

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340.

Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59(1):24–37.

Duchêne, D. A., Bragg, J. G., Duchêne, S., Neaves, L. E., Potter, S., Moritz, C., Johnson, R. N., Ho, S. Y. W., and Eldridge, M. D. B. (2017). Analysis of phylogenomic tree space resolves relationships among marsupial families. *Systematic Biology*, 67(3):400–412.

Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, 63(1):1–19.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94:447–462.

Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., Bryant, D., Steel, M. A., Lockhart, P. J., Penny, D., and Martin, W. (2004). A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution*, 21(9):1643–1660.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410.

Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y.-C., Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich, S. J., Hahn, M. W., and Besansky, N. J. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217):1258524.

Gatesy, J. and Springer, M. S. (2014). Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution*, 80:231–266.

Gillespie, J. H. and Langley, C. H. (1979). Are evolutionary rates really variable? *Journal of Molecular Evolution*, 13:27–34.

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70:7–17.

Hein, J., Schierup, M., and Wiuf, C. (2004). *Gene genealogies, variation and evolution: a primer in coalescent theory.* Oxford University Press, USA.

Heled, J. and Drummond, A. J. (2009). Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution*, 27(3):570–580.

Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487:94–98.

Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 167(2):747–760.

Hillis, D. M., Heath, T. A., and St John, K. (2005). Analysis and visualization of tree space. *Systematic Biology*, 54(3):471–482.

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein-sequence data. *Evolution*, 37(1):203–217.

Inagaki, Y., Susko, E., Fast, N. M., and Roger, A. J. (2004). Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1α phylogenies. *Molecular Biology and Evolution*, 21(7):1340–1349.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M. V., Alfaro-Núñez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jønsson, K. A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics*, 22(4):225–231.

Kearns, A. M., Restani, M., Szabo, I., Schrøder-Nielsen, A., Kim, J. A., Richardson, H. M., Marzluff, J. M., Fleischer, R. C., Johnsen, A., and Omland, K. E. (2018). Genomic evidence of speciation reversal in ravens. *Nature Communications*, 9(1):906.

Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24.

Kück, P., Mayer, C., Wägele, J.-W., and Misof, B. (2012). Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS ONE*, 7(5):e36593.

Lanier, H. C. and Knowles, L. L. (2012). Is recombination a problem for species-tree analyses? *Systematic Biology*, 61:691–701.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Leigh, J. W., Lapointe, F.-J., Lopez, P., and Bapteste, E. (2011). Evaluating phylogenetic congruence in the post-genomic era. *Genome Biology and Evolution*, 3:571–587.

Liu, K. J., Dai, J., Truong, K., Song, Y., Kohn, M. H., and Nakhleh, L. (2014a). An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology*, 10:e1003649.

Liu, L. and Edwards, S. V. (2009). Phylogenetic analysis in the anomaly zone. *Systematic Biology*, 58:452–460.

Liu, L., Xi, Z., and Davis, C. C. (2014b). Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. *Molecular Biology and Evolution*, 32(3):791–805.

Liu, L., Yu, L., and Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

Mallet, J., Besansky, N., and Hahn, M. W. (2016). How reticulated are species? *BioEssays*, 38(2):140–149.

Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D., and Lefeuvre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics*, 26(19):2462–2463.

Mendes, F. K. and Hahn, M. W. (2018). Why concatenation fails near the anomaly zone. *Systematic Biology*, 67(1):158–169.

Mendes, F. K., Livera, A. P., and Hahn, M. W. (2019). The perils of intralocus recombination for inferences of molecular convergence. *Philosophical Transactions of the Royal Society B*, 374:20180244.

Mirarab, S., Bayzid, M. S., and Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65(3):366–80.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Novikova, P. Y., Hohmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G., Guggisberg, A., Paape, T., Schmid, K., Fedorenko, O. M., Holm, S., Sall, T., Schlotterer, C., Marhold, K., Widmer, A., Sese, J., Shimizu, K. K., Weigel, D., Kramer, U., Koch, M. A., and Nordborg, M. (2016). Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics*, 48(9):1077–1082.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*, 34(8):2101–2114.

Page, R. D. (1996). On consensus, confidence, and "total evidence". *Cladistics*, 12(1):83–92.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5):568–583.

Pease, J., Haak, D., Hahn, M. W., and Moyle, L. C. (2016). Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology*, 14:e1002379.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*, 9(3):e1000602.

Philippe, H. and Roure, B. (2011). Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC Biology*, 9(1):91.

Philippe, H., Vienne, D. M. d., Ranwez, V., Roure, B., Baurain, D., and Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, 283:1–25.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*, 5(1):50.

Pollard, D. A., Iyer, V. N., Moses, A. M., and Eisen, M. B. (2006). Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genetics*, 2(10):e173.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Rannala, B. and Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology*, page syw119.

Richards, E. J., Brown, J. M., Barley, A. J., Chong, R. A., and Thomson, R. C. (2018). Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological? *Systematic biology*, 67(5):847–860.

Roch, S., Nute, M., and Warnow, T. (2018). Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. *Systematic Biology*, 68(2):281–297.

Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62.

Rogers, J., Muthuswamy, R., Harris, R. A., Mailund, T., Leppälä, K., Athanasiadis, G., Schierup, M. H., Cheng, J., Munch, K., Walker, J. A., Konkel, M. K., Jordan, V., Steely, C. J., Beckstrom, T. O., Bergey, C., Burrell, A., Schrempf, D., Noll, A., Kothe, M., Kopp, G. H., Liu, Y., Murali, S., Billis, K., Martin, F. J., Muffato, M., Cox, L., Else, J., Disotell, T., Muzny, D. M., Phillips-Conroy, J., Aken, B., Eichler, E. E., Marques-Bonet, T., Kosiol, C., Batzer, M. A., Hahn, M. W., Tung, J., Zinner, D., Roos, C., Jolly, C. J., Gibbs, R. A., Worley, K. C., and the Baboon Genome Analysis Consortium (2019). The comparative genomics and complex population history of *Papio* baboons. *Science Advances*, 5:eaau6947.

Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology*, 61(2):225–247.

Scornavacca, C. and Galtier, N. (2017). Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66:112–120.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16:472–482.

Spence, J. P., Kamm, J. A., and Song, Y. S. (2016). The site frequency spectrum for general coalescents. *Genetics*, 202(4):1549–1561.

Springer, M. S. and Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94:1–33.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Lewis, P. O., and Rogers, J. S. (2001). Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology*, 50(4):525–539.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.

Takahata, N. (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, 122(4):957–966.

Takahata, N. and Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, 110(2):325–344.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26(2):119–164.

Vachaspati, P. and Warnow, T. (2015). ASTRID: Accurate species trees from internode distances. *BMC Genomics*, 16(10):S3.

Wakeley, J. (2009). Coalescent theory. *Roberts & Company, Greenwood Village, Colorado*.

Wang, H.-C., Susko, E., and Roger, A. J. (2019). The relative importance of modeling site pattern heterogeneity versus partition-wise heterotachy in phylogenomic inference. *Systematic Biology*, 68(6):1003–1019.

Warnow, T. (2015). Concatenation analyses in the presence of incomplete lineage sorting. *PLoS Currents*, page doi: 10.1371/currents.tol.8d41ac0f13d1abedf4c4a59f5d17b1f7.

Wen, D., Yu, Y., Zhu, J., and Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, 67(4):735–740.

White, D. J., Bryant, D., and Gemmell, N. J. (2013). How good are indirect tests at detecting recombination in human mtDNA? *G3: Genes, Genomes, Genetics*, 3(7):1095–1104.

Wu, M., Kostyun, J. L., Hahn, M. W., and Moyle, L. C. (2018). Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. *Molecular Ecology*, 27:3301–3316.

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2018). Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution*, 35(2):504–517.

Chapter 4.1 Phylogenomics and Genome Annotation

Anamaria Necsulea
Laboratoire de Biométrie et Biologie Évolutive
UMR 5558, CNRS, Université de Lyon, Université Lyon 1
Villeurbanne, France
anamaria.necsulea@univ-lyon1.fr
https://orcid.org/0000-0001-9861-7698

Abstract

Annotating a genome is a challenging endeavor, which aims to describe not only the protein-coding and non-coding gene catalogues, but also other functional elements involved in gene expression regulation, maintenance of genome integrity and genome transmission across generations. Recent technical developments have greatly improved the annotation process by providing large-scale assessments of transcription, translation, chromatin status and tri-dimensional conformation etc... Genome-wide maps of various biochemical activities can thus be readily obtained. However, biochemical activity is not synonymous with biological function and many active genomic elements may in fact be dispensable. Genome editing techniques allow for more direct tests of biological functions, but are still costly, time-consuming, and largely limited to phenotypes that can be observed in the laboratory. In this context, evolutionary approaches, which can identify genomic regions under purifying selection to preserve existing functions, or under positive selection following the acquisition of new biological roles, are an important asset for functional genome annotation. While evolutionary analyses cannot determine precise biological functions, they can be used to test for functionality at multiple levels, by assessing selective pressures on primary DNA or RNA sequences, on secondary RNA structures, transcription levels or patterns, transcription factor binding sites etc... Here, I review the proven and potential contributions of phylogenomic approaches to genome annotation, focusing on how these methods can be combined with insights from molecular biology and genetics to provide a comprehensive image of functional genomic landscapes.

How to cite: Anamaria Necsulea (2020). Phylogenomics and Genome Annotation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 4.1, pp. 4.1:1–4.1:26. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Introduction

Understanding how complex biological functions are encoded in the DNA is a fundamental goal of genetics. An important step towards attaining this goal is the process of genome annotation, which aims to describe the localization, structure, biochemical activities and (ideally) biological roles of the functional elements present in a genome.

The scope of genome annotation has expanded in recent years. When the first complete DNA sequences of cellular organisms were obtained (Fleischmann et al., 1995), annotating a genome was largely synonymous with describing its catalogue of protein-coding genes. This endeavor is challenging in itself, as demonstrated by the fact that, almost twenty years after the initial publication of the human genome sequence (Lander et al., 2001), the number of protein-coding genes present in our genome has yet to reach a stable estimate (Pertea et al.,

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 4.1; pp. 4.1:1–4.1:26

A book completely handled by researchers.
No publisher has been paid.

2018b). For eukaryotes, annotating protein-coding genes is complicated by the presence of complex exon-intron structures and of multiple isoforms for each gene. Expectedly, alternative transcript annotations are even less stable than known gene repertoires, as thousands of new isoforms are added at each genome annotation release for human or mouse (Harrow et al., 2012). Thus, annotating the complete protein-coding gene repertoire is in itself an ambitious aim.

More recently, describing non-coding RNA genes has become an important part of the genome annotation process. Some categories of non-coding RNAs, such as ribosomal or transfer RNAs, which have essential roles in translating messenger RNAs (mRNAs) into proteins, have been extensively studied and are thus generally well annotated in most species (Abe et al., 2014). Other classes of non-coding RNAs are more elusive. These include both small RNAs (such as miRNAs, which regulate gene expression at the post-transcriptional and translational level (He and Hannon, 2004), or piRNAs, which are thought to protect the germline from transposable element invasion (Weick and Miska, 2014)) and large RNAs (such as long non-coding RNAs, which were proposed to act in a multitude of biological processes [Guttman et al. 2009]). In vertebrates, the number of annotated non-coding RNA genes has increased exponentially in the past few years, thanks to the development of sensitive transcriptome sequencing techniques (Wang et al., 2009). For example, the human genome may harbor as many as 60,000 long non-coding RNA (lncRNA) genes (Iyer et al., 2015; Pertea et al., 2018a), which vastly surpasses the number of known protein-coding genes.

Efforts to chart the functional components of a genome now go even beyond establishing a complete protein-coding and non-coding gene list. In addition to gene repertoires, comprehensive genome annotation projects aim to survey elements that are important for gene expression regulation, for the maintenance of genome integrity, genome transmission across generations, etc. . . Such integrative functional annotation projects are in progress for the human and mouse genomes (Carninci et al., 2005; ENCODE Project Consortium et al., 2007), as well as for other model organisms (modENCODE Consortium et al., 2010; Gerstein et al., 2010). These aspects of genome annotation were made possible by technological advances that enabled large-scale surveys of various biochemical activities, such as enhancer activity (Visel et al., 2009), transcription factor binding (Robertson et al., 2007) or initiation of DNA replication (Cadoret et al., 2008).

Regardless of the class of genomic element that is annotated, either genic or non-genic, biological function is far more difficult to assess than biochemical activity (see Chapter 4.2 [Robinson-Rechavi 2020]). Indeed, numerous genomic elements are biochemically active but functionally dispensable (Graur et al., 2015). For example, transcriptional activity is often observed for pseudogenes, long after the loss of biological functions (Nakamura et al., 2009). A direct test for functionality is to examine the phenotypes and fitness of individuals in which specific elements are inactivated through genetic manipulations. Until recently, genetic manipulation techniques could only be applied to a few targeted genomic loci at a time and were exclusively used with laboratory-grown model organisms or to cell cultures (Hérault et al., 1998; Hockemeyer et al., 2011; Barde et al., 2011). With the development of CRISPR/Cas-based gene editing techniques (Jinek et al., 2012), these approaches have become more broadly applicable, leading to functional surveys encompassing thousands of loci at a time (Shalem et al., 2014; Sanjana et al., 2016). However, at the moment these techniques are still costly, time-consuming and largely restricted to phenotypes that can be observed in the laboratory. In this context, evolutionary studies can bring important insights into the functionality of diverse genomic elements. While precise biological functions generally cannot be predicted through evolutionary analyses, they are useful tools for predicting genome

functionality, by revealing elements that have been under purifying selection to preserve existing biological roles, or under positive selection following the acquisition of new functions.

Here, I review the contributions of large-scale evolutionary genomic (or phylogenomic) approaches to genome annotation. Focusing on eukaryotes, I will present several aspects of genome annotation, such as delineating the protein-coding and non-coding gene repertoires, describing gene expression regulatory elements and identifying other functional genomic elements or structures. I will present the molecular biology and genetic techniques that are nowadays frequently employed to generate data for genome annotation, as well as the evolutionary approaches that can be used to bring insights into the functionality of various genetic elements. I will thus endeavor to show how these methods can be combined to provide a comprehensive image of functional genomic landscapes.

2 Annotating protein-coding and non-coding gene repertoires

Undoubtedly the most important step of the genome annotation process is to characterize gene repertoires. This is a complex procedure, which can be roughly divided into four steps: describing gene models, predicting broad functional categories of genes (e.g., protein-coding and non-coding genes), inferring gene functionality and annotating putative gene functions. Here, I will discuss how phylogenomic approaches can contribute to these four gene annotation steps.

2.1 Gene model description

Describing gene models in eukaryotes is a challenging task, which involves identifying transcribed regions, transcription start and end sites, exon-intron structures and alternative splicing variants. Gene model prediction can be performed either *ab initio*, using species-specific data and predictive methods, or through homology-based approaches, which use gene and protein information from closely-related species to predict genes in the species of interest. *Ab initio* methods are evidently required for organisms where genome sequences for closely-related species are lacking. Conversely, homology-based predictions are beneficial when data from closely-related species is abundant, and were notably used to annotate primate genomes (Chimpanzee Sequencing and Analysis Consortium, 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium et al., 2007).

For *ab initio* gene model prediction, transcriptome sequencing has become an invaluable tool. In its many forms, transcriptome sequencing has long benefited genome annotation efforts, even before next-generation sequencing techniques became available. For example, analyses of expressed sequence tags (ESTs) helped compile the initial catalogue of human genes (Lander et al., 2001), and Cap Analysis of Gene Expression (CAGE) sequencing data were used to annotate mouse gene promoters (Carninci et al., 2005). More recently, massively parallel transcriptome sequencing methods (commonly termed RNA-seq), have become an indispensable aspect of the genome annotation process. Compared to previous transcriptomics assays, RNA-seq offers increased sequencing depth and thus higher transcript detection sensitivity, even for moderately expressed genes (Wang et al., 2009). To improve detection sensitivity even at low expression levels, RNA-seq can be used in combination with RT-PCR amplification (Howald et al., 2012) or with capture on tiling arrays (Clark et al., 2015; Bussotti et al., 2016), which considerably increases the sequencing depth for targeted transcripts or genomic regions. Several computational methods were developed to assemble transcript sequences from RNA-seq data, either using a genome sequence as a reference (Trapnell et al., 2010; Pertea et al., 2015) or entirely *de novo* (Grabherr et al., 2011). The

application of transcriptome sequencing to genome annotation has revealed many forms of transcriptome complexity. These include the presence of numerous transcript variants for protein-coding genes, generated through canonical mechanisms such as alternative splicing, use of alternative transcription initiation or termination sites, read-through transcription or trans-splicing (ENCODE Project Consortium et al., 2007; Gerstein et al., 2007). These in-depth genome annotation studies also established that transcription is pervasive outside of protein-coding genes (ENCODE Project Consortium et al., 2007). In particular, in-depth transcriptome and chromatin accessibility surveys revealed that mammalian genomes contain tens of thousands of long non-coding RNAs (Guttman et al., 2009; Khalil et al., 2009; Iyer et al., 2015; Pertea et al., 2018b).

Homology-based gene model prediction approaches are of particular importance for non-model species, when other sources of data are insufficient. The quality of a genome annotation largely depends on the quality and quantity of transcriptomic and proteomic data available for that species (Mudge and Harrow, 2016). For widely-studied species such as human, mouse, fruitfly or nematode, extensive resources (including full-length or partial cDNA sequences, RNA-seq and proteomics data) have accumulated over time and are available as input for genome annotation (Mudge and Harrow, 2016). However, this is an exception rather than the rule, and for many species experimental data are scarce. In this case, homology-based annotation methods can be applied, with relative facility. The most frequently used gene model prediction software, including Augustus (Stanke et al., 2006), Gnomon (Souvorov et al., 2010), Exonerate (Slater and Birney, 2005) and GeneWise (Birney et al., 2004), can use as input protein and RNA sequences from closely related species. In the simplest implementations, the genome is scanned to identify local alignments between protein-sequences and nucleotide sequence translations. This is for example done in the Ensembl annotation pipeline (Zerbino et al., 2018), in which pairwise alignments between reference protein sequences and translated nucleotide sequences are generated and exploited to predict gene structures, with Exonerate (Slater and Birney, 2005) and GeneWise (Birney et al., 2004). The efficiency of this approach depends on the degree of sequence conservation between the proteins used as reference and the ones encoded in the target genome. To identify more divergent proteins, an extension of Augustus (Keller et al., 2011) uses multiple sequence alignments to construct protein conservation profiles and to identify blocks of ungapped, highly-conserved sequences. Predicted gene structures in the target genome are then compared with the resulting sequence conservation profiles, and are assigned higher confidence scores if they match the amino acid composition profiles of conserved alignment blocks.

Homology-based prediction methods can also be insightful for annotating non-coding RNA genes. For lncRNAs, which are generally weakly expressed, defining gene models with standard RNA-seq data is often not sufficient, as the low read coverage can result in gene model fragmentation (Howald et al., 2012). In these cases, for comparative analyses of lncRNAs across closely-related species, it can be beneficial to project annotations from one species to another, based on primary sequence similarity (Washietl et al., 2014; Necsulea et al., 2014). This method has obvious disadvantages, as it cannot correctly analyze homologous lncRNA loci that have diverged in terms of exon/intron structures, nor can it predict loci where transcription is species-specific (Hezroni et al., 2015). Homology-based annotation approaches, for protein-coding genes, non-coding RNAs or other types of functional genomic elements, all share these limitations, and it is important to complement these methods with species-specific "omics" data. Nevertheless, they provide a valuable starting point on which more comprehensive genome annotation resources can be built.

2.2 Gene classification

A second important step in the genome annotation process, after gene model description, is to provide a broad classification of the resulting loci into protein-coding and non-coding genes. This step is more difficult than it can seem at first sight, mainly because lncRNAs are structurally very similar to protein-coding mRNAs (Derrien et al., 2012).

To categorize genes as protein-coding or non-coding, direct proteome assays are an evident path. However, proteomics technologies, although in continuous progress (Richards et al., 2015), are still far from the throughput observed for RNA-seq. Large-scale investigations of the proteome based on mass spectrometry have only recently become available for humans (Kim et al., 2014; Wilhelm et al., 2014), and are still lacking for most other species. Recent studies were able to detect and quantify peptides for approximately 84% of annotated protein-coding genes, but generally lacked power to detect known alternative protein isoforms (Kim et al., 2014; Wilhelm et al., 2014). In the absence of high-throughput proteome sequencing, an alternative avenue towards large-scale investigations of the proteome (or at least of the translatome) is provided by the development of ribosome profiling (Ingolia et al., 2009). This technique isolates and sequences RNA molecules that are bound by poly-ribosome complexes, which are thus likely actively translated (Ingolia et al., 2009). While more accessible than mass spectrometry, this technique is nevertheless considerably more complex than classical RNA-seq, and very little data has been generated so far. Thus, experimental data that could help distinguish between protein-coding and non-coding RNA genes are not readily available. Instead, computational methods, many of which are based on the patterns of sequence evolution, have been developed to determine the protein-coding potential of newly-annotated transcripts.

It is interesting to note that the first long non-coding RNA ever identified in mammals, namely the H19 lncRNA, was defined as such using an evolutionary approach (Brannan et al., 1990). Sequence analyses of the mouse transcript revealed the presence of several small open reading frames (ORFs). However, comparisons with the human homolog showed that none of these open reading frames were conserved during evolution, indicating that the locus did not encode a functional protein (Brannan et al., 1990). Indeed, the mere presence of ORFs is not a reliable indicator that an eukaryotic sequence is protein-coding, given that such stretches can appear by chance in long RNA molecules (Clamp et al., 2007). In contrast, their conservation during evolution, through negative selection that prevents the fixation of ORF-disrupting mutations, is a strong predictor of the presence of a constrained protein-coding sequence. The idea of exploiting the patterns of sequence evolution to predict the protein-coding potential of genomic sequences was later implemented into two computational methods that aimed to detect *bona fide* protein-coding genes in yeast and fruitfly genomes: the reading frame conservation (RFC) method (Kellis et al., 2004) and the codon substitution frequency (CSF) method (Lin et al., 2007). The RFC method assesses the presence of ORF-disrupting insertions and deletions in a multiple sequence alignment between the target species and other "informant" species (Kellis et al., 2004). The CSF method (Figure 1) analyzes the proportion of synonymous and non-synonymous single-nucleotide substitutions between the target and informant species, in all possible reading frames (Lin et al., 2007). Given that it relies on the presence of insertions and deletions, which are less frequent than point mutations, the RFC approach strongly depends on the degree of sequence conservation between the target and informant species (Lin et al., 2008). In contrast, the CSF method has high sensitivity and specificity values, although it may propose wrong classifications for protein-coding sequences that are subject to positive selection (Lin et al., 2008, 2011). This approach was used to distinguish protein-coding and non-coding regions in the first

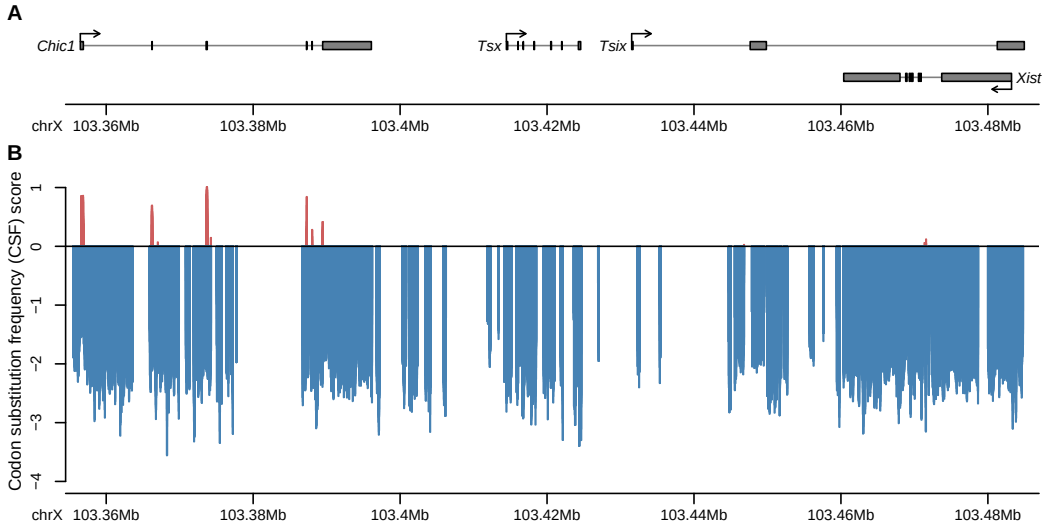


Figure 1 The codon substitution frequency (CSF) score exploits the pattern of nucleotide substitutions in a multiple species alignment to predict protein-coding regions. A) Genomic localization and exon-intron structure for mouse *Chic1*, *Tsx*, *Tsix* and *Xist* genes. The rectangles represent the exons and the arrows represent the direction of transcription. B) The codon substitution frequency (CSF) score variation in the same genomic region. Positive CSF scores, which indicate the presence of protein-coding regions under purifying selection to preserve protein sequences, mainly co-localize with *Chic1* annotated protein-coding exons. Another annotated protein-coding gene, *Tsx*, does not show any positive scores. Negative CSF scores are observed elsewhere, including on the exons of long non-coding RNAs *Xist* and *Tsix*. Whole-genome CSF data were taken from a previous publication (Necsulea et al., 2014); recently, whole-genome PhyloCSF data have become available (Mudge et al., 2019).

large-scale investigations of lncRNAs (Guttman et al., 2009; Khalil et al., 2009) and later in the first large-scale evolutionary analyses of lncRNA across vertebrates (Necsulea et al., 2014). Although its efficiency is higher for longer sequences, if sufficient "informants" are included in the analysis (including both distant and closely related species with respect to the species of interest), the CSF method can also detect short protein-coding regions. This approach can thus be applied to scan protein-coding regions in the whole genome, with a sliding window approach (Figure 1, Mudge et al., 2019).

Expectedly, gene classifications as protein-coding or non-coding obtained with biochemical or evolutionary approaches do not always agree. Notably, ribosome profiling studies revealed that numerous lncRNAs annotated with evolutionary approaches are in fact actively translated (Ingolia et al., 2014), and mass spectrometry assays were able to detect peptide sequences stemming from hundreds of lncRNAs (Kim et al., 2014). While some of this inconsistency could simply be attributed to imperfect sensitivity and specificity of the classification methods, the presence of ribosome footprints on lncRNA sequences is not itself evidence that these transcripts are translated into functional proteins. In fact, the ribosome occupancy profile is strikingly different between genuine protein-coding mRNAs and lncRNAs: while a sharp ribosome release at the stop codon and a strong reading frame preference is observed for the former, the profiles are much more uniform along lncRNA sequences, indicating that these transcripts are simply scanned by ribosomes, but likely do not generate functional proteins (Guttman et al., 2013).

Another intriguing cause of disagreement between the phylogenomic and biochemical

classification methods is the evolutionary history of the genes (see Chapter 4.2 [Robinson-Rechavi 2020]). Indeed, the RFC and CSF methods both rely on the pattern of sequence evolution, which can be assessed within varying evolutionary time frames, depending on the phylogenetic relatedness of the analyzed species. However, the functional category of the gene may itself evolve over time. For example, protein-coding genes may become pseudogenized, and potentially resurrected into functional lncRNAs, as is famously the case for *Xist* (Duret et al., 2006), as well as for other conserved lncRNAs (Hezroni et al., 2017). Conversely, lncRNAs may transform into protein-coding genes by acquiring functional ORFs (McLysaght and Hurst, 2016). This evolutionary plasticity highlights the importance of combining phylogenomic and biochemical approaches to determine the protein-coding potential of newly annotated transcripts, which may reveal insights into the evolutionary processes that lead to new gene origination (McLysaght and Hurst, 2016).

2.3 Gene functionality

The staggering complexity of the human transcriptome (Pertea et al., 2018b; Iyer et al., 2015; Carninci et al., 2005) raises the question of its functionality. Many of the transcripts discovered with high-throughput transcriptome sequencing data, whether alternative isoforms of protein-coding genes, read-through transcripts that join neighboring genes and in particular long non-coding RNAs, may in fact be functionally dispensible, representing so-called "transcriptional noise" (Ponjavic et al., 2007). Experimental methods that directly address functionality typically rely on genetic manipulations that inactivate or over-express specific transcripts, followed by phenotypic evaluations. Although these methods have recently become more accessible, applicable to large numbers of loci (Shalem et al., 2014; Joung et al., 2017) and to a wider range of organisms (Mazo-Vargas et al., 2017), they are still costly, time-consuming and largely restricted to phenotypes that can be observed in the laboratory. In this context, phylogenomic approaches are extremely valuable, as they can provide solid predictions of biological functionality (Haerty and Ponting, 2014).

The ongoing search for lncRNA functionality is a good illustration of the usefulness of phylogenomic methods in this context. Indeed, in the absence of large-scale experimental data for this category of genes, the functionality of lncRNAs has often been investigated with evolutionary approaches. One such study compared the rates and patterns of sequence evolution between mammalian long non-coding RNAs and ancient transposable element insertions, which are likely neutrally-evolving (Ponjavic et al., 2007). This study revealed slightly, but significantly lower rates of evolution for lncRNAs than for ancient repeats, indicating the presence of purifying selection for at least a subset of lncRNAs (Ponjavic et al., 2007). These conclusions were confirmed by subsequent studies, which consistently showed that mammalian lncRNAs are more conserved than expected by chance, but that they display modest levels of primary sequence conservation compared to protein-coding genes (Guttman et al., 2009; Washietl et al., 2014; Necsulea et al., 2014; Marques and Ponting, 2009; Kutter et al., 2012; Haerty and Ponting, 2013; Wiberg et al., 2015). These studies assessed either long-term selective constraints, for example by analyzing PhastCons scores determined from whole-genome alignments of placental mammals or vertebrates (Figure 2), or short-term sequence evolution, contrasting single-nucleotide polymorphisms within populations and sequence divergence between closely related species (Haerty and Ponting, 2013; Wiberg et al., 2015).

In contrast, in fruitfly, lncRNAs are under strong purifying selection (Haerty and Ponting, 2013; Young et al., 2012). These observations are in agreement with the "transcriptional noise" hypothesis, and the differences between mammals and fruitfly likely reflect the reduced

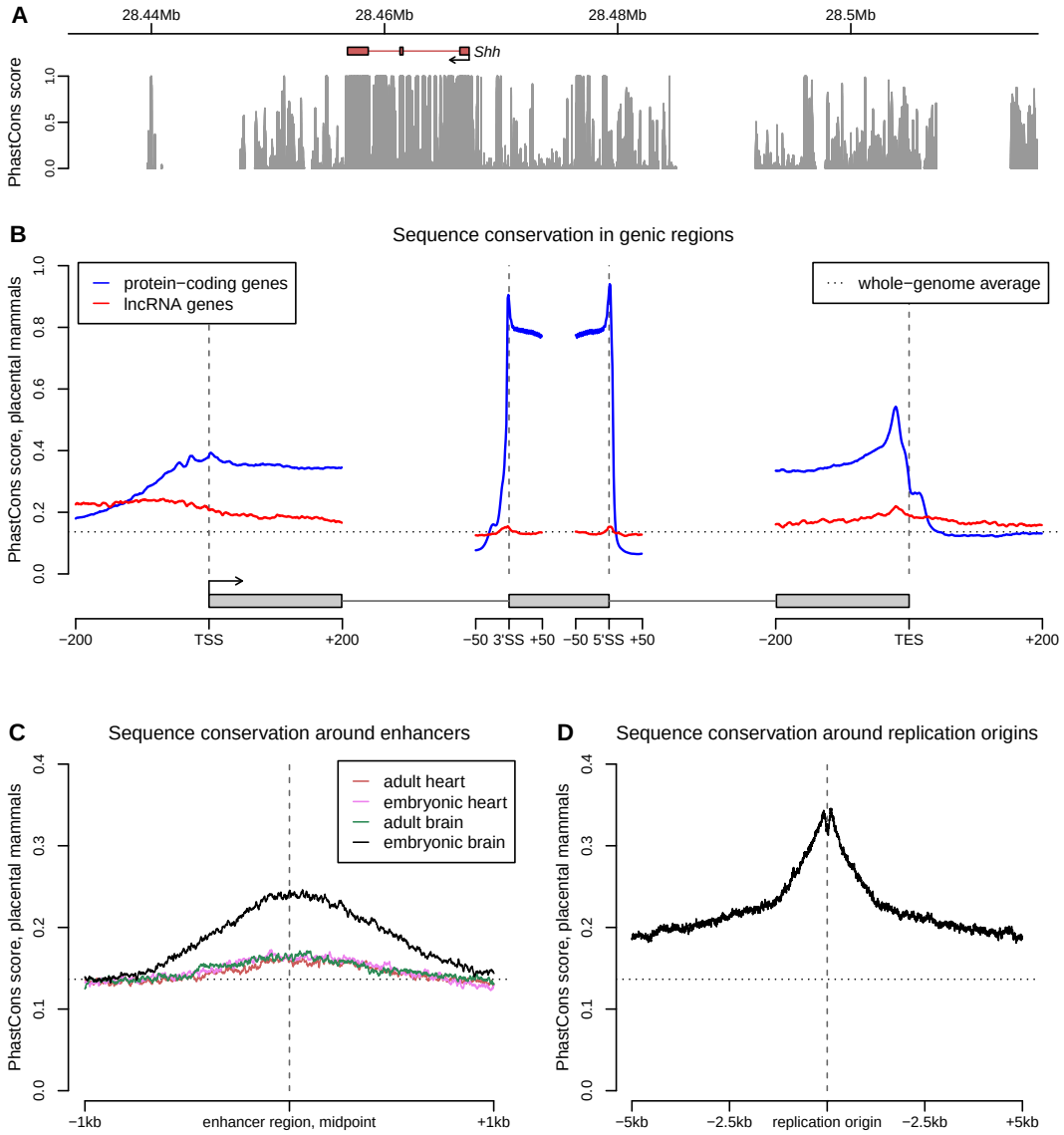


Figure 2 Sequence conservation patterns around functional elements in the mouse genome. A) Sequence conservation (PhastCons score (Siepel et al., 2005), computed on a whole-genome alignment of mouse and 59 other vertebrate species) variation around the *Shh* gene, in the mouse genome. The amount of sequence conservation reaches maximum values in *Shh* exons, but also in neighboring intergenic regions, potentially including regulatory elements. B) Average sequence conservation profile in protein-coding and lncRNA gene structures: transcription start sites, splice sites and transcription end sites. C) Average sequence conservation profiles around mouse transcriptional enhancers (Yue et al., 2014) from different tissues. D) Average sequence conservation profiles around mouse replication origins (Cayrou et al., 2015). B-D) The average sequence conservation profiles were based on the PhastCons score, computed on a whole-genome alignment of mouse and 39 other placental mammal species (Siepel et al., 2005). PhastCons scores were downloaded from the UCSC Genome Browser (Casper et al., 2018).

efficiency of natural selection in the former, due to low effective population sizes (Haerty and Ponting, 2013).

However, alternative hypotheses were proposed to explain the low levels of sequence constraint observed for mammalian lncRNAs without dismissing their potential functionality. A plausible hypothesis posits that lncRNA functions may be achieved by short sequence motifs, which may for example mediate their binding to genomic regions or protein sequences (Hezroni et al., 2015). This would explain why levels of evolutionary conservation, when computed on the entire length of lncRNAs, are only slighty above neutral expectations (Ponjavic et al., 2007). Interestingly, analyses of human lncRNAs revealed that almost all sequence constraint is indeed concentrated in very short sequence motifs, but that these small constrained regions are in fact splicing regulatory elements (Figure 2; Schüler et al., 2014; Haerty and Ponting, 2015). Purifying selection on sequences needed to achieve correct splicing of multi-exonic lncRNA loci could indeed be indicative of transcript functionality. However, a recent experimental investigation showed that splicing of lncRNA loci can influence the expression of neighboring genes (Engreitz et al., 2016). Thus, the presence of selection on lncRNA splicing motifs does not necessarily prove that lncRNA transcripts are themselves biologically functional.

Another hypothesis that could explain the weak levels of lncRNA conservation is that selective pressures may act on secondary RNA structures, rather than on primary transcript sequences (Kapusta and Feschotte, 2014). This hypothesis can be directly tested, for example by contrasting the degree of RNA secondary structure conservation with the degree of primary sequence conservation, using RNA structures predicted with thermodynamic modeling and multiple sequence alignments (Washietl et al., 2005). Using this principle, genome-wide scans for conserved RNA secondary structures consistently confirmed selective pressures on miRNA, tRNA and rRNA structures (Figure 3), but revealed only limited such constraint within long non-coding RNA loci (Pedersen et al., 2006; Parker et al., 2011; Seemann et al., 2017).

Overall, there is increasing evidence that lncRNA functionality often does not reside in the RNA molecule encoded by the locus, but in the presence of additional regulatory elements that affect neighboring gene expression patterns (Latos et al., 2012; Engreitz et al., 2016; Amândio et al., 2016). Experimental studies of lncRNA functions must be carefully designed to address these strong confounding effects (Bassett et al., 2014). Likewise, phylogenomic studies of lncRNA functionality need to be adapted to account for additional targets of selective pressures (Haerty and Ponting, 2014).

2.4 Gene function

Even when gene models (*i.e.*, gene localization, exon-intron structure and alternative isoforms) can be predicted based on species-specific experimental data, gene functions are still overwhelmingly inferred based on homology. Indeed, experimental investigations of protein or RNA functions are lagging well behind the vast amounts of transcripts and proteins predicted from next-generation sequencing data. Functional annotations are thus commonly transferred across species based on homology relationships, with the underlying assumption that gene functions are generally conserved during evolution (see Chapter 4.2 [Robinson-Rechavi 2020]). As for homology-based gene model predictions, the efficacy and reliability of the transfer of functional annotations across species is dependent on the degree of sequence divergence between the reference sequences and the target genome to be annotated. Computational methods that can predict homologous gene families in the presence of high degrees of sequence divergence are thus of great interest (Vilella et al., 2009). Another important challenge is to correctly identify gene duplication events, and to predict the functional characteristics of the resulting gene copies. Indeed, gene duplication is believed to be an important driver of

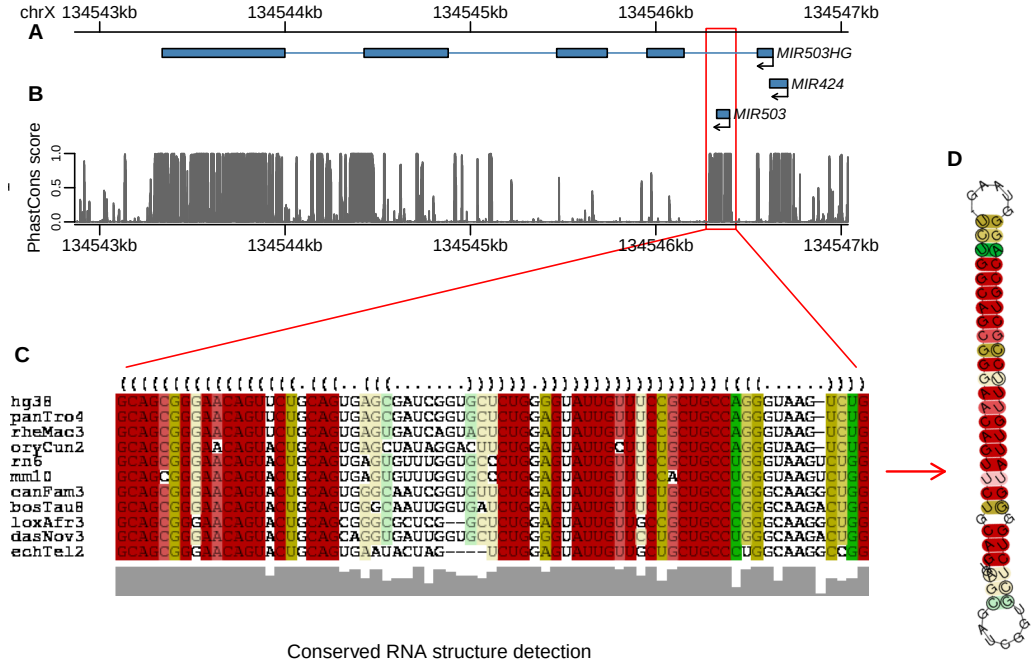


Figure 3 Identification of conserved RNA structures using the pattern of sequence evolution (Seemann et al., 2017). A) Genomic position and exon-intron structure for lncRNA gene *MIR503-HG* and miRNA genes *MIR503* and *MIR424*, in the human genome. The rectangles represent the exons and the arrows represent the direction of transcription. B) Sequence conservation profile (PhastCons score [Siepel et al. 2005], computed on a whole-genome alignment of human and 99 other vertebrate genomes), on the same genomic region. PhastCons scores were provided by the UCSC Genome Browser (Casper et al., 2018). C) Sequence alignment and predicted consensus RNA structure in the *MIR503* region. D) Resulting conserved RNA structure for *MIR503*.

functional innovation, as the initially redundant gene copies can accumulate mutations that lead to sub-functionalization or to neo-functionalization (Conant and Wolfe, 2008). For both homologous and paralogous genes, the likelihood of functional conservation decreases with increasing divergence time (Studer and Robinson-Rechavi, 2009). The relationship between the extent of sequence (or structure) divergence and functional divergence cannot be readily defined, and it likely varies among functional categories of genes (Tian and Skolnick, 2003). Thus, cross-species projections of gene functions need to be interpreted with great caution.

Homology-based gene model annotation and functional assignment methods have been applied to both protein-coding and non-coding genes. However, these approaches are significantly more successful for the former than for the latter, as non-coding RNA sequences are generally much less conserved than protein sequences. Among non-coding RNA classes, lncRNAs in particular evolve very rapidly (Figure 2; Washietl et al., 2014; Necsulea et al., 2014). This is well illustrated by the fact that lncRNA annotation efforts based on gene model projections across species could identify only approximately 2,000 lncRNAs conserved in placental mammals (Washietl et al., 2014; Necsulea et al., 2014). These studies predicted conserved lncRNAs based on primary sequence conservation and required species-specific transcription evidence to confirm the activity of the lncRNA loci in other species (Washietl et al., 2014; Necsulea et al., 2014). Here again, additional methodological developments are needed to exploit the specific patterns of lncRNA evolution, such as the presence of

short stretches of conserved regions within larger, overall divergent sequences (Hezroni et al., 2015). Transfer of functional annotations across species is particularly problematic for long non-coding RNAs, for which experimental data on biological functions are scarce even in model organisms. In this context, comparative transcriptomics analysis across species can provide crude functional assignments, for example by identifying evolutionarily conserved co-expression relationships between lncRNAs and protein-coding genes, which may indicate functional associations (Stuart et al., 2003; Necsulea et al., 2014).

3 Annotating non-genic functional elements with phylogenomic approaches

Eukaryotic genomes harbor numerous functional non-genic elements. These include non-coding sequences that regulate gene expression, such as transcriptional enhancers (Banerji et al., 1981) or silencers (Busturia et al., 1997), splicing regulatory elements (Lee and Rio, 2015), but also origins of DNA replication (Benbow et al., 1992), insulators that organize chromatin architecture in the nucleus (Van Bortle and Corces, 2012), recombination hotspots (Smith, 1994), etc... Some categories of non-coding functional elements can be now be identified with dedicated experimental assays, such as chromatin immunoprecipitation and sequencing (ChIP-seq) techniques that identify genomic sequences bound by specific proteins or by modified histones (Robertson et al., 2007; Visel et al., 2009), or nascent DNA strand sequencing to pinpoint origins of replication (Cadoret et al., 2008; Cayrou et al., 2015). However, by construction these techniques use the presence of biochemical activity to predict biological function, although the two concepts are far from being synonymous (Graur et al., 2013). Indeed, numerous biochemically active genomic elements are altogether dispensable from a biological point of view, either because most cellular mechanisms (including transcription, protein-DNA binding, etc...) are error-prone, or because of functional redundancy with other genomic elements (Graur et al., 2013). Additional data are thus needed to ascertain biological functionality, and phylogenomic approaches are again a valuable asset in this context.

Perhaps the most striking example of how phylogenomic approaches can be used to annotate functional non-coding elements is the discovery of ultra-conserved sequences (Duret et al., 1993; Bejerano et al., 2004). These elements were first identified through comparative analyses of nucleotide sequences across distant vertebrate species, which revealed the presence of regions with unexpectedly high degrees of conservation (more than 70% sequence similarity for species that diverged at least 300 million years ago, Duret et al., 1993). This pioneering study, which predates the genomic era, was later confirmed through genome-wide scans, which identified thousands of ultra-conserved elements outside of protein-coding genes in vertebrates and in other metazoan genomes (Bejerano et al., 2004; Siepel et al., 2005). Importantly, the low rate of sequence evolution in these regions is not due to overlap with mutational cold-spots. On the contrary, analyses of within-species polymorphism and between-species divergence rates showed that these elements are subject to intense purifying selective pressures (Katzman et al., 2007), which further underscores their functional relevance. *In vivo* experimental assays showed that a great proportion of ultraconserved elements have transcriptional enhancer capacity in the mouse embryo (Pennacchio et al., 2006), thus confirming the regulatory roles proposed upon their initial discovery (Duret et al., 1993). Some of these elements may also belong to non-coding RNA loci (Kern et al., 2015).

It is important to stress that phylogenomic approaches that focus on signatures of strong evolutionary conservation cannot discover all types of functional non-coding elements.

For example, the extreme levels of sequence conservation observed for some embryonic transcriptional enhancers are not observed in all tissues and developmental stages: heart enhancers show much weaker levels of sequence conservation than brain enhancers (Blow et al., 2010), and enhancers active in adult brain are much less conserved than those active in embryonic brain (Figure 2). Other functional genomic elements, such as origins of replication, also display increased sequence conservation compared to the genomic background (Figure 2, Cadoret et al., 2008). However, much of the sequence conservation observed within experimentally predicted origins of DNA replication in the human genome stems from their overlap with transcriptional promoters (Cadoret et al., 2008).

In addition to overlooking genomic elements that are under weak purifying selection, which are difficult to distinguish from the neutrally evolving genomic background, phylogenomic scans may also bypass functional elements that evolve rapidly due to positive selection. Dedicated computational methods were developed to identify genomic regions that evolve faster than expected under a neutral regime (Pollard et al., 2010). However, an accelerated rate of sequence evolution, which is the main signal used to predict the footprints of adaptation in non-coding regions, is by no means synonymous with positive selection. Biased gene conversion, a non-adaptive mechanism that promotes the fixation of specific alleles in highly recombining regions, frequently leads to accelerated sequence evolution, thereby confounding positive selection scans (Duret and Galtier, 2009; Ratnakumar et al., 2010).

Phylogenomic approaches that aim to predict functional non-genic elements will likely further be improved by the increasing numbers of complete genome sequences, including population genomics datasets that enable investigations of DNA sequence variations within and between populations (1000 Genomes Project Consortium et al., 2015), in addition to between-species sequence divergence. Moreover, important efforts have been made to generate combined genome and transcriptome population datasets, such as Geuvadis (Lappalainen et al., 2013) or GTEX (GTEx Consortium, 2015). Joint analyses of genome and transcriptome variations within populations have already been used to predict putative regulatory variants, that is, polymorphisms that are statistically associated with expression level variations between individuals (Lappalainen et al., 2013; GTEx Consortium, 2015). Combined with between-species genome and transcriptome comparative analyses, these approaches could bring insights into the selective pressures that act on gene expression levels (Gilad et al., 2006; Romero et al., 2012), and thereby help annotate non-coding RNA transcripts whose expression patterns are constrained, rather than their RNA sequences (Latos et al., 2012).

4 Combining molecular biology, genetics and evolutionary biology to annotate functional genomic elements

We have never been this close to truly uncovering the functional landscapes of the genomes. In the past decade, technological innovations have enabled us not only to investigate biochemical activities (such as transcription, translation or transcription factor binding) at a genome-wide level, but also to perform large-scale experimental assessments of biological functions through genetic manipulations (Jinek et al., 2012; Sanjana et al., 2016; Joung et al., 2017). The contributions of molecular biology and genetics methodologies to functional genome annotation are thus indisputable. However, even in this technology-dominated context, phylogenomic approaches are still an invaluable tool for the discovery and annotation of functional genomic elements.

Phylogenomic methods, such as genome-wide scans for regions under purifying or positive selection, can be used in combination with molecular biology assays and genetic manipulations

to obtain thorough functional characterizations for specific genomic elements. First of all, very often, genetic manipulation studies use the presence of evolutionary sequence conservation to prioritize elements for further experiments (Sauvageau et al., 2013). Moreover, evolutionary analyses can also provide information into the facet of a locus that is most likely the target of natural selection, and which should thus be perturbed through genetic manipulations to test for biological function. For example, for long non-coding RNAs the highest degrees of sequence conservation were observed on promoter regions and splicing regulatory elements (Figure 2, Guttman et al., 2009; Ponjavic et al., 2007; Schüler et al., 2014; Haerty and Ponting, 2015). Genetic manipulations later showed that the presence of transcription and splicing at multiple lncRNA loci affected neighboring gene expression, while the production of a specific RNA sequence was dispensable (Engreitz et al., 2016). Thus, the functional elements in lncRNA loci could be correctly predicted with an evolutionary approach.

While most phylogenomic studies can bring insights into the functionality of a given locus (that is, on its effect on the overall fitness of the organism), rather than on its specific biological functions, in some cases evolutionary studies can go even beyond and predict the mode of action or the phenotype in which a genomic element is involved. For example, genome-wide scans for evolutionarily conserved RNA secondary structures have uncovered thousands of genomic regions that are transcribed into structured non-coding RNAs, such as miRNAs, tRNAs or rRNAs (Pedersen et al., 2006; Parker et al., 2011; Seemann et al., 2017). Interestingly, while most phylogenomic scans for functional elements relie on the presence of evolutionary conservation, evolutionary losses of genes and other genomic elements can also bring insights into genomic functions. An elegant evolutionary approach aiming to discover genes and regulatory elements that are involved in specific phenotypes is the recently proposed "forward genomics" method, which analyzes phylogenies in which the same specific trait (e.g. the ability to synthesize vitamin C) was lost multiple times independently (Hiller et al., 2012). Genomic regions that were needed only to achieve the specific function under study are likely to accumulate substitutions in the lineages that have lost it, due to relaxation of purifying selection pressures. This approach can successfully predict genes and non-genic functional elements that are specifically associated with a given trait, if sufficient independent trait losses can be analyzed (Hiller et al., 2012). Althouh this methodology clearly has limitations, not least of which is the pervasive presence of pleiotropy in vertebrate genomes, it is an exciting use of phylogenomics for functional genome annotation, which bridges the gap between genome and phenotypes.

So far, phylogenomic methods have been successfully used to predict gene localization and structure, expression regulatory elements, conserved RNA secondary structures, as well as to distinguish between coding and non-coding transcribed regions. As molecular biology and genetic technologies continue to progress, bringing us closer to understanding genomic functions, the field of evolutionary genomics must also continue to develop and to propose new methods to assess selective pressures that act on newly discovered classes of functional elements. We can thus hope to make sense of the intricate functional architecture of our genomes, in the light of evolution (Haerty and Ponting, 2014).

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Abe, T., Inokuchi, H., Yamada, Y., Muto, A., Iwasaki, Y., and Ikemura, T. (2014). tRNADB-CE: tRNA gene database well-timed in the era of big sequence data. *Front Genet*, 5:114.

Amândio, A. R., Necsulea, A., Joye, E., Mascrez, B., and Duboule, D. (2016). Hotair is dispensible for mouse development. *PLoS Genet.*, 12(12):e1006232.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308.

Barde, I., Verp, S., Offner, S., and Trono, D. (2011). Lentiviral Vector Mediated Transgenesis. *Curr Protoc Mouse Biol*, 1(1):169–184.

Bassett, A. R., Akhtar, A., Barlow, D. P., Bird, A. P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A. C., Gingeras, T. R., Haerty, W., Higgs, D. R., Miska, E. A., and Ponting, C. P. (2014). Considerations when investigating lncRNA function in vivo. *Elife*, 3:e03058.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325.

Benbow, R. M., Zhao, J., and Larson, D. D. (1992). On the nature of origins of DNA replication in eukaryotes. *Bioessays*, 14(10):661–670.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res*, 14(5):988–995.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Bristow, J., Ren, B., Black, B. L., Rubin, E. M., Visel, A., and Pennacchio, L. A. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.*, 42(9):806–810.

Brannan, C. I., Dees, E. C., Ingram, R. S., and Tilghman, S. M. (1990). The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.*, 10(1):28–36.

Bussotti, G., Leonardi, T., Clark, M. B., Mercer, T. R., Crawford, J., Malquori, L., Notredame, C., Dinger, M. E., Mattick, J. S., and Enright, A. J. (2016). Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome Res.*, 26(5):705–716.

Busturia, A., Wightman, C. D., and Sakonju, S. (1997). A silencer is required for maintenance of transcriptional repression throughout Drosophila development. *Development*, 124(21):4343–4350.

Cadoret, J.-C., Meisch, F., Hassan-Zadeh, V., Luyten, I., Guillet, C., Duret, L., Quesneville, H., and Prioleau, M.-N. (2008). Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl. Acad. Sci. U.S.A.*, 105(41):15837–15842.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S.,

McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. a. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563.

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Karolchik, D., Hinrichs, A. S., Haeussler, M., Guruvadoo, L., Navarro Gonzalez, J., Gibson, D., Fiddes, I. T., Eisenhart, C., Diekhans, M., Clawson, H., Barber, G. P., Armstrong, J., Haussler, D., Kuhn, R. M., and Kent, W. J. (2018). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.*, 46(D1):D762–D769.

Cayrou, C., Ballester, B., Peiffer, I., Fenouil, R., Coulombe, P., Andrau, J.-C., van Helden, J., and Méchali, M. (2015). The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res.*, 25(12):1873–1885.

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, 104(49):19428–19433.

Clark, M. B., Mercer, T. R., Bussotti, G., Leonardi, T., Haynes, K. R., Crawford, J., Brunck, M. E., Cao, K.-A. L., Thomas, G. P., Chen, W. Y., Taft, R. J., Nielsen, L. K., Enright, A. J., Mattick, J. S., and Dinger, M. E. (2015). Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat Methods*, 12(4):339–342.

Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.*, 9(12):938–950.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J., and Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.*, 22(9):1775–1789.

Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*, 312(5780):1653–5.

Duret, L., Dorkeld, F., and Gautier, C. (1993). Strong conservation of non-coding sequences during vertebrates evolution: Potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res*, 21(10):2315–2322.

Duret, L. and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*, 10:285–311.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern,

A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., and Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, 539(7629):452–455.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, 17(6):669–681.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dosé, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D. M., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Rätsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., modENCODE Consortium, Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., and Waterston, R. H. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, 330(6012):1775–1787.

Gilad, Y., Oshlack, A., and Rifkin, S. A. (2006). Natural selection on gene expression. *Trends Genet.*, 22(8):456–461.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N.,

and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–652.

Graur, D., Zheng, Y., and Azevedo, R. B. R. (2015). An evolutionary classification of genomic function. *Genome Biol Evol*, 7(3):642–645.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol*, 5(3):578–590.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235):223–227.

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., and Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, 154(1):240–251.

Haerty, W. and Ponting, C. P. (2013). Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.*, 14(5):R49.

Haerty, W. and Ponting, C. P. (2014). No gene in the genome makes sense except in the light of evolution. *Annu Rev Genomics Hum Genet*, 15:71–92.

Haerty, W. and Ponting, C. P. (2015). Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA*, 21(3):333–346.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.*, 22(9):1760–1774.

He, L. and Hannon, G. J. (2004). MicroRNAs: Small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, 5(7):522–531.

Hérault, Y., Rassoulzadegan, M., Cuzin, F., and Duboule, D. (1998). Engineering chromosomes in mice through targeted meiotic recombination (TAMERE). *Nat. Genet.*, 20(4):381–384.

Hezroni, H., Ben-Tov Perry, R., Meir, Z., Housman, G., Lubelsky, Y., and Ulitsky, I. (2017). A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biology*, 18(1):1–15.

Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*, 11(7):1110–1122.

Hiller, M., Schaar, B. T., Indjeian, V. B., Kingsley, D. M., Hagey, L. R., and Bejerano, G. (2012). A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep*, 2(4):817–823.

Hockemeyer, D., Wang, H., Kiani, S., Lai, C. S., Gao, Q., Cassady, J. P., Cost, G. J., Zhang, L., Santiago, Y., Miller, J. C., Zeitler, B., Cherone, J. M., Meng, X., Hinkley, S. J., Rebar,

E. J., Gregory, P. D., Urnov, F. D., and Jaenisch, R. (2011). Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.*, 29(8):731–734.

Howald, C., Tanzer, A., Chrast, J., Kokocinski, F., Derrien, T., Walters, N., Gonzalez, J. M., Frankish, A., Aken, B. L., Hourlier, T., Vogel, J.-H., White, S., Searle, S., Harrow, J., Hubbard, T. J., Guigó, R., and Reymond, A. (2012). Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.*, 22(9):1698–1710.

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., Wills, M. R., and Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep*, 8(5):1365–1379.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223.

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y.-M., Robinson, D. R., Beer, D. G., Feng, F. Y., Iyer, H. K., and Chinnaiyan, A. M. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*, 47(3):199–208.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821.

Joung, J., Engreitz, J. M., Konermann, S., Abudayyeh, O. O., Verdine, V. K., Aguet, F., Gootenberg, J. S., Sanjana, N. E., Wright, J. B., Fulco, C. P., Tseng, Y.-Y., Yoon, C. H., Boehm, J. S., Lander, E. S., and Zhang, F. (2017). Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature*, 548(7667):343–346.

Kapusta, A. and Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: Mechanisms and biological implications. *Trends Genet.*, 30(10):439–452.

Katzman, S., Kern, A. D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R. K., Salama, S. R., and Haussler, D. (2007). Human genome ultraconserved elements are ultraselected. *Science*, 317(5840):915.

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, 27(6):757–763.

Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E. S. (2004). Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *J. Comput. Biol.*, 11(2-3):319–355.

Kern, A. D., Barbash, D. A., Chang Mell, J., Hupalo, D., and Jensen, A. (2015). Highly constrained intergenic Drosophila ultraconserved elements are candidate ncRNAs. *Genome Biol Evol*, 7(3):689–698.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S., and Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 106(28):11667–11672.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram,

S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H., and Pandey, A. (2014). A draft map of the human proteome. *Nature*, 509(7502):575–581.

Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., Odom, D. T., and Marques, A. C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.*, 8(7):e1002841.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de

Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Geuvadis Consortium, Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Häsler, R., Syvänen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigó, R., Gut, I. G., Estivill, X., and Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

Latos, P. A., Pauler, F. M., Koerner, M. V., Şenergin, H. B., Hudson, Q. J., Stocsits, R. R., Allhoff, W., Stricker, S. H., Klement, R. M., Warczok, K. E., Aumayr, K., Pasierbek, P., and Barlow, D. P. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science*, 338(6113):1469–1472.

Lee, Y. and Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.*, 84:291–323.

Lin, M. F., Carlson, J. W., Crosby, M. A., Matthews, B. B., Yu, C., Park, S., Wan, K. H., Schroeder, A. J., Gramates, L. S., St Pierre, S. E., Roark, M., Wiley, K. L., Kulathinal, R. J., Zhang, P., Myrick, K. V., Antone, J. V., Celniker, S. E., Gelbart, W. M., and Kellis, M. (2007). Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res.*, 17(12):1823–1836.

Lin, M. F., Deoras, A. N., Rasmussen, M. D., and Kellis, M. (2008). Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput. Biol.*, 4(4):e1000067.

Lin, M. F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–282.

Marques, A. C. and Ponting, C. P. (2009). Catalogues of mammalian long noncoding RNAs: Modest conservation and incompleteness. *Genome Biol.*, 10(11):R124.

Mazo-Vargas, A., Concha, C., Livraghi, L., Massardo, D., Wallbank, R. W. R., Zhang, L., Papador, J. D., Martinez-Najera, D., Jiggins, C. D., Kronforst, M. R., Breuker, C. J., Reed, R. D., Patel, N. H., McMillan, W. O., and Martin, A. (2017). Macroevolutionary shifts of WntA function potentiate butterfly wing-pattern diversity. *Proc. Natl. Acad. Sci. U.S.A.*, 114(40):10701–10706.

McLysaght, A. and Hurst, L. D. (2016). Open questions in the study of de novo genes: What, how and why. *Nat. Rev. Genet.*, 17(9):567–578.

modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., Washietl, S., Arshinoff, B. I., Ay, F., Meyer, P. E., Robine, N., Washington, N. L., Di Stefano, L., Berezikov, E., Brown, C. D., Candeias, R., Carlson, J. W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M. Y., Will, S., Alekseyenko, A. A., Artieri, C., Booth, B. W., Brooks, A. N., Dai, Q., Davis, C. A., Duff, M. O., Feng, X., Gorchakov, A. A., Gu, T., Henikoff, J. G., Kapranov, P., Li, R., MacAlpine, H. K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S. K., Riddle, N. C., Sakai, A., Samsonova, A., Sandler, J. E., Schwartz, Y. B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K. H.,

Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S. E., Brent, M. R., Cherbas, L., Elgin, S. C. R., Gingeras, T. R., Grossman, R., Hoskins, R. A., Kaufman, T. C., Kent, W., Kuroda, M. I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J. W., Ren, B., Russell, S., Cherbas, P., Graveley, B. R., Lewis, S., Micklem, G., Oliver, B., Park, P. J., Celniker, S. E., Henikoff, S., Karpen, G. H., Lai, E. C., MacAlpine, D. M., Stein, L. D., White, K. P., and Kellis, M. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330(6012):1787–1797.

Mudge, J. M. and Harrow, J. (2016). The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.*, 17(12):758–772.

Mudge, J. M., Jungreis, I., Hunt, T., Gonzalez, J. M., Wright, J. C., Kay, M., Davidson, C., Fitzgerald, S., Seal, R., Tweedie, S., He, L., Waterhouse, R. M., Li, Y., Bruford, E., Choudhary, J. S., Frankish, A., and Kellis, M. (2019). Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Res.*, 29(12):2073–2087.

Nakamura, K., Akama, T., Bang, P. D., Sekimura, S., Tanigawa, K., Wu, H., Kawashima, A., Hayashi, M., Suzuki, K., and Ishii, N. (2009). Detection of RNA expression from pseudogenes and non-coding genomic regions of Mycobacterium leprae. *Microb. Pathog.*, 47(3):183–187.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Grutzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505(7485):635–640.

Parker, B. J., Moltke, I., Roth, A., Washietl, S., Wen, J., Kellis, M., Breaker, R., and Pedersen, J. S. (2011). New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res*, 21(11):1929–43.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comp Biol*, 2(4):e33.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, 33(3):290–295.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y.-C., Madugundu, A. K., Pandey, A., and Salzberg, S. L. (2018a). CHESS: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.*, 19(1):208.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Chang, Y.-C., Madugundu, A. K., Pandey, A., and Salzberg, S. (2018b). Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *Biorxiv*.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1):110–121.

Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*, 17(5):556–65.

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M. T. (2010). Detecting positive selection within genomes: The problem of biased gene conversion. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 365(1552):2571–2580.

Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., Remington, K. A., Strausberg, R. L., Venter, J. C., Wilson, R. K., Batzer, M. A., Bustamante, C. D., Eichler, E. E., Hahn, M. W., Hardison, R. C., Makova, K. D., Miller, W., Milosavljevic, A., Palermo, R. E., Siepel, A., Sikela, J. M., Attaway, T., Bell, S., Bernard, K. E., Buhay, C. J., Chandrabose, M. N., Dao, M., Davis, C., Delehaunty, K. D., Ding, Y., Dinh, H. H., Dugan-Rocha, S., Fulton, L. A., Gabisi, R. A., Garner, T. T., Godfrey, J., Hawes, A. C., Hernandez, J., Hines, S., Holder, M., Hume, J., Jhangiani, S. N., Joshi, V., Khan, Z. M., Kirkness, E. F., Cree, A., Fowler, R. G., Lee, S., Lewis, L. R., Li, Z., Liu, Y.-S., Moore, S. M., Muzny, D., Nazareth, L. V., Ngo, D. N., Okwuonu, G. O., Pai, G., Parker, D., Paul, H. A., Pfannkoch, C., Pohl, C. S., Rogers, Y.-H., Ruiz, S. J., Sabo, A., Santibanez, J., Schneider, B. W., Smith, S. M., Sodergren, E., Svatek, A. F., Utterback, T. R., Vattathil, S., Warren, W., White, C. S., Chinwalla, A. T., Feng, Y., Halpern, A. L., Hillier, L. W., Huang, X., Minx, P., Nelson, J. O., Pepin, K. H., Qin, X., Sutton, G. G., Venter, E., Walenz, B. P., Wallis, J. W., Worley, K. C., Yang, S.-P., Jones, S. M., Marra, M. A., Rocchi, M., Schein, J. E., Baertsch, R., Clarke, L., Csürös, M., Glasscock, J., Harris, R. A., Havlak, P., Jackson, A. R., Jiang, H., Liu, Y., Messina, D. N., Shen, Y., Song, H. X.-Z., Wylie, T., Zhang, L., Birney, E., Han, K., Konkel, M. K., Lee, J., Smit, A. F. A., Ullmer, B., Wang, H., Xing, J., Burhans, R., Cheng, Z., Karro, J. E., Ma, J., Raney, B., She, X., Cox, M. J., Demuth, J. P., Dumas, L. J., Han, S.-G., Hopkins, J., Karimpour-Fard, A., Kim, Y. H., Pollack, J. R., Vinar, T., Addo-Quaye, C., Degenhardt, J., Denby, A., Hubisz, M. J., Indap, A., Kosiol, C., Lahn, B. T., Lawson, H. A., Marklein, A., Nielsen, R., Vallender, E. J., Clark, A. G., Ferguson, B., Hernandez, R. D., Hirani, K., Kehrer-Sawatzki, H., Kolb, J., Patil, S., Pu, L.-L., Ren, Y., Smith, D. G., Wheeler, D. A., Schenck, I., Ball, E. V., Chen, R., Cooper, D. N., Giardine, B., Hsu, F., Kent, W. J., Lesk, A., Nelson, D. L., O'brien, W. E., Prüfer, K., Stenson, P. D., Wallace, J. C., Ke, H., Liu, X.-M., Wang, P., Xiang, A. P., Yang, F., Barber, G. P., Haussler, D., Karolchik, D., Kern, A. D., Kuhn, R. M., Smith, K. E., and Zwieg, A. S. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316(5822):222–234.

Richards, A. L., Merrill, A. E., and Coon, J. J. (2015). Proteome sequencing goes deep. *Curr Opin Chem Biol*, 24:11–17.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4(8):651–657.

Robinson-Rechavi, M. (2020). Molecular evolution and gene function. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.2, pages 4.2:1–4.2:20. No commercial publisher | Authors open access book.

Romero, I. G., Ruvinsky, I., and Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.*, 13(7):505–516.

Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., and Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. *Science*, 353(6307):1545–1549.

Sauvageau, M., Goff, L. A., Lodato, S., Bonev, B., Groff, A. F., Gerhardinger, C., Sanchez-

Gomez, D. B., Hacisuleyman, E., Li, E., Spence, M., Liapis, S. C., Mallard, W., Morse, M., Swerdel, M. R., D'Ecclessis, M. F., Moore, J. C., Lai, V., Gong, G., Yancopoulos, G. D., Frendewey, D., Kellis, M., Hart, R. P., Valenzuela, D. M., Arlotta, P., and Rinn, J. L. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife*, 2:e01749.

Schüler, A., Ghanbarian, A. T., and Hurst, L. D. (2014). Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.*, 31(12):3164–3183.

Seemann, S. E., Mirza, A. H., Hansen, C., Bang-Berthelsen, C. H., Garde, C., Christensen-Dalsgaard, M., Torarinsson, E., Yao, Z., Workman, C. T., Pociot, F., Nielsen, H., Tommerup, N., Ruzzo, W. L., and Gorodkin, J. (2017). The identification and functional annotation of RNA structures conserved in vertebrates. *Genome Res.*, 27(8):1371–1383.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., and Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166):84–87.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–50.

Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31.

Smith, G. R. (1994). Hotspots of homologous recombination. *Experientia*, 50(3):234–241.

Souvorov, A., Kapustin, Y., Kiryutin, B., Chetvernin, V., Tatusova, T., and Lipman, D. (2010). Gnomon – NCBI eukaryotic gene prediction tool. *NCBI*.

Stanke, M., Tzvetkova, A., and Morgenstern, B. (2006). AUGUSTUS at EGASP: Using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.*, 7 Suppl 1:S11.1–8.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255.

Studer, R. A. and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.*, 25(5):210–216.

Tian, W. and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, 333(4):863–882.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511–515.

Van Bortle, K. and Corces, V. G. (2012). Nuclear organization and genome function. *Annu. Rev. Cell Dev. Biol.*, 28:163–187.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19(2):327–335.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63.

Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, 102(7):2454–2459.

Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res*, 24:616–28.

Weick, E.-M. and Miska, E. A. (2014). piRNAs: From biogenesis to function. *Development*, 141(18):3458–3471.

Wiberg, R. A. W., Halligan, D. L., Ness, R. W., Necsulea, A., Kaessmann, H., and Keightley, P. D. (2015). Assessing Recent Selection and Functionality at Long Noncoding RNA Loci in the Mouse Genome. *Genome Biol Evol*, 7(8):2432–2444.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F., and Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587.

Young, R. S., Marques, A. C., Tibbit, C., Haerty, W., Bassett, A. R., Liu, J.-L., and Ponting, C. P. (2012). Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. *Genome Biol Evol*, 4(4):427–442.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., Kent, W. J., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Hansen, R. S., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, M. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A., Ren, B., and Mouse ENCODE Consortium (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515(7527):355–364.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken,

B. L., Cunningham, F., Yates, A., and Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Res.*, 46(D1):D754–D761.

Chapter 4.2 Molecular Evolution and Gene Function

Marc Robinson-Rechavi
Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland
Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Abstract

One of the basic questions of phylogenomics is how gene function evolves, whether among species or inside gene families. In this chapter, we provide a brief overview of the problems associated with defining gene function in a manner which allows comparisons which are both large scale and evolutionarily relevant. The main source of functional data, despite its limitations, is transcriptomics. Functional data provides information on evolutionary mechanisms primarily by showing which functional classes of genes evolve under stronger or weaker purifying or adaptive selection, and on which classes of mutations (e.g., substitutions or duplications). However, the example of the "ortholog conjecture" shows that we are still not at a point where we can confidently study phylogenomically the evolution of gene function at a precise scale.

How to cite: Marc Robinson-Rechavi (2020). Molecular Evolution and Gene Function. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 4.2, pp. 4.2:1–4.2:20. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 The problem with "function"

Molecular evolution interacts with gene function in two fundamental ways. First, different gene families will evolve differently according to their function, e.g. they are under different selection pressures on their protein sequence or on their diversification by gene duplication. Second, gene function itself evolves. Both of these assertions are quite obvious in their generality. Problems arise when we try to characterize more specific patterns, and to test more specific hypotheses. While no aspect of phylogenomics is without its difficulties, this is a particularly vexing one: what is gene function? Two distinctions are fundamental to the study of function. First, between healthy and pathological function, i.e. what the gene does when it is present and functional, *versus* what is disrupted when the gene is absent or somehow not functioning properly. The latter includes most medical genetics observations, as well as Knock-Out/Knock-Down phenotypes. Second, we need to distinguish between selected effect and causal role. This second distinction has been abundantly discussed following the publication of ENCODE 2012 (Pennisi, 2012; The ENCODE Project Consortium, 2012; Doolittle, 2013; Eddy, 2013; Graur et al., 2013; Germain et al., 2014; Graur et al., 2015). ENCODE is a large collaborative project to "build a comprehensive parts list of functional elements in the human genome", based on systematic biochemical assays, such as RNA-seq or ChIP-seq, in different cell types. The observation that ≈80% of the human genome had some type of biochemical activity in some cell type led to statements that all that DNA was functional (Pennisi, 2012; The ENCODE Project Consortium, 2012). The questions of function and of evolution are tightly linked in biology because it is natural selection which explains the functional adaptation of organisms and their parts (see Chapter 4.1 [Necsulea 2020]). The function of the lungs is to breath, i.e. to exchange oxygen and CO_2 between the


A book completely handled by researchers.
No publisher has been paid.

Evidence gene A	Evidence gene A'	Apparent conclusion	Relevance
Experiment X: function x	Homology transfer: function x	Conserved function	No: circular reasoning
Experiment X: function x	Experiment Y: function y	Different function	No: experiments cannot be compared
Experiment X: function x	Experiment X: function x	Conserved function	Yes: evolutionary conservation
Experiment X: function x	Experiment X: function x'	Different function	Yes: evolutionary change

Table 1 Evidence for function of homologous genes and evolutionary relevance. A and A' are homologous genes.

organism and the air. This comes neither from intention of the lungs nor of the organism, but because ancestors of some vertebrates which were better at exchanging oxygen and CO_2 with the air had better survival and reproductive success. Thus it has been proposed that function be defined as that which a structure was selected to do. This is the "selected-effect definition of function" (Doolittle et al., 2014). The lungs were selected to exchange gases, not to develop cancers or take space in the thoracic cage, although they also do these things. An alternative definition of function, the "causal role" definition, does not appeal to evolutionary history, and could in fact include such features as the lungs taking space, or the nose supporting sunglasses (Doolittle et al., 2014). The same questions and definitions apply to all levels of biological organization, including genes. In the aftermath of ENCODE, much of the focus has been on classifying DNA sequences as "functional" or not. This question is more directly relevant to genome annotation (see Chapter 4.1 [Necsulea 2020]). For this chapter, we will mostly focus on protein coding genes, for which we have strong a priori reasons to expect that they are indeed functional. One simple line of evidence is that genes which are sufficiently conserved among species to undertake phylogenomics studies are most probably conserved by purifying selection, and thus functional. But to understand the role of function in molecular evolution beyond the generality that functional sequences are more conserved, we need to focus on classifying their specific functions. One way to classify specific gene functions is to collect assertions and evidence from the published biological literature (Thomas, 2017). The largest undertaking in this sense is the Gene Ontology consortium (see Box 1.1). The Gene Ontology describes the selected effect function of gene products, whether they are proteins or functional RNAs. Thus it notably does not describe pathological roles, which are typically causal role functions.

From a phylogenomic perspective, the properties of the Gene Ontology and its annotations have important consequences. These annotations can only ever capture knowledge at a given point in time, and they capture it from a disparate collection of studies with differing aims and methods. Thus even genes with evolutionarily conserved functions will often have different annotations, because of different experiments (e.g. Altenhoff et al., 2012; Chen and Zhang, 2012), see Table 1. Moreover many genes are never or very rarely the object of targeted experimental studies (Sinha et al., 2018).

These limitations are not specific of the Gene Ontology, but will affect any effort to capture gene function from the abundance of precise but heterogeneous experimental data. For example, Enzyme Classification (E.C.) numbers (McDonald and Tipton, 2014) have been

Box 1.1: The Gene Ontology

The Gene Ontology is composed of three ontologies, which describe different aspects of gene function (Ashburner et al., 2000; Dessimoz and Skunca, 2016). Briefly, the Cellular Component ontology describes where in or out of a cell the gene product is found; the Molecular Function ontology describes the activity of the gene product, potentially as part of a protein or RNA complex; the Biological Process ontology describes the result of the organismal program in which the gene product acts. As can be readily seen, the latter is more complex than the other two. The Molecular Function can be thought of as "what does the gene product do in a test tube?", while the Biological Process can be thought of as "what does the gene product do within the organism?". Being ontologies, all three include not only standard terms and definitions, but also relations between the terms. These relations form a directed graph, meaning that (i) there is a direction to the relations, for example "steroid binding" is_a "lipid binding" but not the inverse, and (ii) terms can have both several children and several parents, for example "steroid bindin" not only is_a "lipid binding" but also is_a "organic cyclic compound binding" and has_input from "steroid" (parents in the graph), while it has ten children, including "steroid hormone binding" and "vitamin D binding". This graph includes very general terms, such as "binding" or "catalytic activity", and very specific terms, such as "17alpha-hydroxyprogesterone binding" or "estrogen response element binding".
The annotation of genes with the Gene Ontology consists in associating each gene with as many Gene Ontology terms as necessary, which describe the known function of the product(s) of this gene. Association can be based on (i) evidence from hypothesis-driven, small-scale, published studies, which provide the closest to selected effect function; (ii) large scale hypothesis-free experiments (such as ENCODE), which provide "candidate functions" (Thomas, 2017), closer to the causal role functional definition; or (iii) electronic inference, whether simply by "best Blast hit" or more advanced domain modelling or text mining.

used to investigate functional evolution, but E.C. numbers are mostly associated to gene products by homology, at the gene or the domain level, thus creating pseudo-evolutionary patterns in the data. If all proteins with homology to a given enzyme obtain a certain E.C. number, then that function will appear conserved, whether it is or not (see Table 1). In the GO, the evidence used for assertions of functional annotation are available in a standard code (Giglio et al., 2018), which allows to distinguish conservation of function between homologs with experimental evidence from patterns due to functional annotation transfer between homologs. Directly comparing the phenotypes associated to genes is even more complicated by the differences among experiments and species, see Box 1.2. A few studies have shown promise in that phenotypes can effectively be compared between distant species (McGary et al., 2010; Kachroo et al., 2015), but the complexity of phenotypes still limits applications such as comparing subtle changes between orthologs or paralogs (see Chapter 2.4 [Fernández et al. 2020] for definitions), or relating functional change to protein evolutionary rates.

An alternative approach to investigate specific gene function is to use genome-wide experiments. While such data have been criticized for biasing GO annotations towards the types of function that can thus be investigated (Schnoes et al., 2013), they can provide comparable functional information across genes and species. Transcriptomics is particularly

interesting because techniques are becoming relatively cheap and straightforward to apply to different species, conditions, or individuals, thus providing a direct link between gene activity and evolution. Yet there are also limitations of these data. Gene expression does not provide information on most aspects of gene function. Transcriptomics informs on (i) where and when a gene is expressed, (ii) how highly it is expressed, and (iii) which genes are co-expressed, but gives little information about which components of the phenotype are involved. On the other hand, transcriptomics provides a direct link between phylogenomics and Evo-Devo, where expression patterns are the main form of evidence.

Box 1.2: Phenotypes and function

Within the selected-effect definition of function, an ideal measure of function would be to relate genes to organismal level phenotypes. But to use them in phylogenomic studies, we need to define and measure phenotypes in a way that is systematic and robust enough.

One basic measure of phenotype impact is essentiality: is loss of a gene lethal to the organism – often extended in sexual organisms to include sterility (Hurst and Smith, 1999; He and Zhang, 2006; Liao and Zhang, 2007; Makino et al., 2009)? While this seems straightforward, the same gene loss can be lethal or not depending on growth conditions (Ooi et al., 2006) or genetic background (Ayadi et al., 2012). This limits the evolutionary interpretation of such results, since natural selection has been acting on genes in a variety of backgrounds and environments.

In unicellular cultivated organisms, such as many bacteria or yeasts, one standardised measure of phenotype for comparisons among paralogs or strains is growth rate in a controlled environment (Hillenmeyer et al., 2008). One positive aspect of such measures is that they are probably closely related to fitness, but on the other hand, they only convey a very unspecific characterization of gene function. To study phenotypes beyond essentiality at a genomic scale between species, they need to be encoded in a standard manner. One promising solution is to develop inter-species phenotype ontologies (Mungall et al., 2010; Robinson et al., 2014; Mungall et al., 2017), but this approach is still limited by the difficulties of annotating phenotypes in different species. A recent study measured growth phenotypes in 32 bacterial species over different conditions (Price et al., 2018). This still only covers a small part of the genes of these species, but it shows promise in the possibility of scaling up to full phylogenomic studies. However, this approach remains restricted to easily cultivated microorganisms.

Finally, two caveats affect almost all measures of phenotype from gene Knock-Out experiments. First, the conditions under which natural selection has acted are expected to be very different from the typical laboratory settings (e.g. Ruff et al., 2015). Secondly, "knocking out" a gene can be done in different ways (complete or partial, conditional or not), and it is not obvious which of these correspond to mutations which could occur in nature and be subject to natural selection. For example comparing phenotypes of essentiality between human and mouse means comparing diverse experimental designs to diverse spontaneous mutations (Liao and Zhang, 2008), or using essentiality in human cell culture.

From a phylogenomic perspective, while it is relatively straightforward to compare gene expression results between paralogs within a species, comparisons between species are more complicated (discussed in Roux et al., 2015). Indeed, the direct comparison of expression

levels is complicated by batch effects (Gilad and Mizrahi-Man, 2015), different organisms being often studied independently. On the other hand, transforming continuous expression values into "expressed" versus "not expressed", which allows comparison between different species and provides a link to Evo-Devo reasoning, loses much of the information from transcriptome data. Correlations of expression levels in different conditions (e.g., different organs) are also problematic (Pereira et al., 2009; Piasecka et al., 2012b). Some of these problems have been evaded by defining qualitative variables summarizing patterns of gene expression, such as tissue specificity, which reflects function while being robust to differences in methods and sampling (Kryuchkova-Mostacci and Robinson-Rechavi 2016, 2017; Chapter 4.3 [Robinson-Rechavi et al. 2020]). An additional complexity of using gene expression in phylogenomics is that samples must be comparable (discussed in Roux et al., 2015). In practice, different organs, developmental stages, sexes, or abiotic conditions can be sampled, and homology or even similarity are not always clear. Even inside one species, for instance when comparing paralogs, care must be taken to distinguish variation in expression across tissues or developmental sequences from changes between experimental, abiotic conditions. Assuming that, despite these many caveats, functional annotation has been achieved in a large enough set of species, one can think about studying the evolution of gene function. Ideally, we would like to know when function changed, and whether the changes were driven by selection or drift. The main approach to this question is based on Ornstein-Uhlenbeck models, which are notably used in the phylogenetic study of gene expression (Bedford and Hartl, 2009). Briefly, a Brownian model of gene expression change is contrasted to models with different optima in different lineages; if there is significant support for different optima, this can be taken as evidence for changes in gene function. While the principle is very attractive, the limited data that we still have leads to issues of lack of power or of over-fitting (e.g. Ho and Ané, 2014; Cooper et al., 2016), and there are problems with phylogenetic studies of expression when species sampling is small (Dunn et al., 2013). Finally, summarizing the expression of many genes in modules is also attractive because of its relevance to the way genes are expected to function as modules in relation to biological processes. These modules can be computed per species, before evolutionary computations (e.g. Piasecka et al., 2013), or computed across species, allowing to detect conserved expression patterns (e.g. Brawand et al., 2011). The clustering itself can also contain information on gene evolution, for example with transcriptomes of eyes of cave-dwelling and surface crayfish clustering by eye function and not according to the phylogenetic relationships of the species (Stern and Crandall, 2018). These aspects are developed further in Section 3.

2 Gene families with different functions evolve differently

Gene function and evolution can interact in two ways: genes with different functions evolve differently, and the function itself evolves. The first aspect is easier to study, as it is less dependent on the detailed specifics of functional annotation. On the other hand, causality can be difficult to determine, as many features of gene function and evolution are correlated. We will present here some of the main trends, keeping in mind that this is a rapidly changing domain.

2.1 Gene expression and function determine protein evolutionary rates

The sequence of different proteins evolves at very different rates, over at least three orders of magnitude (see Chapters 2.1 and 5.1 [Simion et al. 2020; Pett and Heath 2020]). Efforts to understand the reasons of this variation have been called a "quest for the universals of

protein evolution" (Rocha, 2006). The most intuitive explanation for these differences is that proteins that are more essential to the organism evolve slower, because of stronger negative selection (selection against change). But studies of the statistical determinants of protein evolutionary rates have shown that reality is more complex (Pal et al., 2006). The "'importance" of proteins, as measured notably by the phenotypic effect of knocking the genes out, predicts only a small fraction of variability. Instead, the strongest predictor of protein evolutionary rates, at least in yeast and E. coli, appears to be the level of expression of the corresponding gene (Rocha and Danchin, 2004). Other significant factors, with a smaller contribution, include mutation rates, recombination rates, protein tertiary structure, and protein-protein interactions (Pal et al., 2006, and Box 2.1). In mammals, the relation of protein sequence evolutionary rate with expression level is weaker, and is mostly explained by breadth of expression among tissues (Duret and Mouchiroud, 2000; Gu and Su, 2007; Larracuente et al., 2008; Kryuchkova-Mostacci and Robinson-Rechavi, 2015), and by expression levels in neural tissues (Gu and Su, 2007; Drummond and Wilke, 2008; Kryuchkova-Mostacci and Robinson-Rechavi, 2015). There is also a correlation in mammals, but not in yeasts, between protein sequence evolutionary rate and changes in expression (Warnefors and Kaessmann, 2013). This variation in mean evolutionary rates reflects differences in purifying selection on protein structure and its capacity to carry out its function. Proteins with different functions are also obviously affected differently by such purifying selection, for two reasons: some gene functions are under stronger selection than others, because they impact phenotype more directly or because they are related to phenotypes which are themselves under stronger selection; and some functions are more directly carried by a specific protein sequence, whereas others less so. For example, histone proteins interact with their whole protein sequence with DNA, thus selection affects all the sequence; and the function of chromatin organisation is fundamental to all cells of an organism, and is under very strong selection. As a result, histones have among the lowest sequence evolutionary rates of any proteins. On the other hand, transcription factors such as the Hox genes are also under strong phenotypic selection, as shown by the conservation of the family (Hoegg and Meyer, 2005), its chromosomal organisation and expression patterns, among distant animals (Hrycaj and Wellik, 2016). Yet Hox protein sequences, like those of many other transcription factors, are very lowly conserved outside of the DNA-binding domain (Hueber et al., 2010). The strong purifying selection does not seem to act directly on most of the protein sequence. Thus different functional categories of genes are under different selective regimes concerning their protein sequences. An additional selective pressure on protein evolutionary rates is that in some tissues, or for some functions, errors in protein synthesis or protein variants have a higher chance of producing misfolded proteins which are toxic to the cell. This leads to optimization of gene sequence to minimize translation and folding errors, and greater intolerance to some types of mutations (Drummond and Wilke, 2009, 2008; Singh et al., 2012).

Protein function also affects sequence evolution through variation in the extent and the mode of positive selection. Continuous positive selection over long evolutionary time has mostly been found on genes involved in sexual selection or immune systems (Obbard et al., 2009; Enard et al., 2016), while episodic positive selection has been found in a wider range of functions (Kosiol et al., 2008; Studer et al., 2008; Barreiro and Quintana-Murci, 2010; Daub et al., 2013, 2017; Slodkowicz and Goldman, 2019). Positive selection patterns are also affected by expression, with more adaptation in genes expressed in the germ-line (Salvador-Martínez et al., 2018), and of genes expressed post-embryonically rather than embryonically (Liu and Robinson-Rechavi, 2018; Coronado-Zamora et al., 2019). Such results are of course

Box 2.1: Network definitions of function

Genes rarely act in isolation, but rather as complexes, networks, or pathways. The information on these gene and protein interactions is difficult to measure accurately at a large scale. Metabolic networks or gene regulatory networks typically integrate information from thousands of precise small-scale experiments, only available in a very small number of model species. Metabolic networks are especially useful to study the phylogenomics of unicellular organisms, and notably bacteria, where evolution by gene gain (by horizontal transfer) and loss is important, and can be understood as adding or removing nodes from such networks (Pal et al., 2005; Noda-Garcia et al., 2018). Gene regulatory networks are especially attractive because they provide a link between phylogenomics and Evo-Devo (Davidson and Erwin, 2006), but robust data at a large scale is rare. Protein-protein interaction networks have been published for several model species, but they still sample the tree of life very sparsely. They have been useful in characterizing differences in evolutionary patterns, e.g., between hub and peripheral proteins (Mintseris and Weng, 2005; Wapinski et al., 2007; Presser et al., 2008), but data sampling and quality are so far not sufficient to directly compare homologous proteins and study the evolution of function (Presser et al., 2008).

dependent on the quality of our positive selection predictions, but they show that to understand adaptation in phylogenomics, we need to take into account gene function.

2.2 Duplication and loss: conservative and dynamic functions

The main mechanism by which genes diversify within genomes is duplication (see Chapters 2.4, 3.1 and 3.2 [Fernández et al. 2020; Schrempf and Szöllosi 2020; Boussau and Scornavacca 2020]). Different molecular mechanisms, such as non-homologous crossover, or transposition, can lead to a DNA region containing one or more genes to be in two or more copies in one haploid genome. Hybridization or abnormal meiosis can lead to polyploidy, in which an individual has extra copies of the whole genome. It is important to keep in mind that these events are mutations. Thus they follow the same dynamics and forces as all mutations. They can rise to fixation in a population or not, under a combination of selection and drift. When polyploidy rises to fixation, and the paralogous copies start diverging, it is often called whole genome duplication (Wolfe, 2001). From the perspective of the evolution of gene function, whole genome duplication and small-scale duplication have important differences (see Figure 1). A whole genome duplication means that duplication of all genes goes to fixation without any impact of the function of each gene. It also means that each gene is duplicated with its full genomic environment, including promoters and enhancers, and that stoichiometry between all gene products is maintained. Conversely, after small-scale duplication, the fixation of the individual duplicated gene will be affected by selection on that gene's function. And duplicate genes can be unequal "at birth" (Kaessmann et al., 2009), if one copy lacks some regulatory elements due to a partial duplication. In all cases, after fixation, duplicate genes can be retained or not. Duplicates are not retained if one copy suffers a nonsense mutation and becomes a pseudogene, and is then eliminated from the genome. If both copies are kept, they can keep the same function or diverge in function, see Figure 1.

As small-scale duplication is much more common (according to some estimates [Lynch

4.2:8 Molecular Evolution and Gene Function

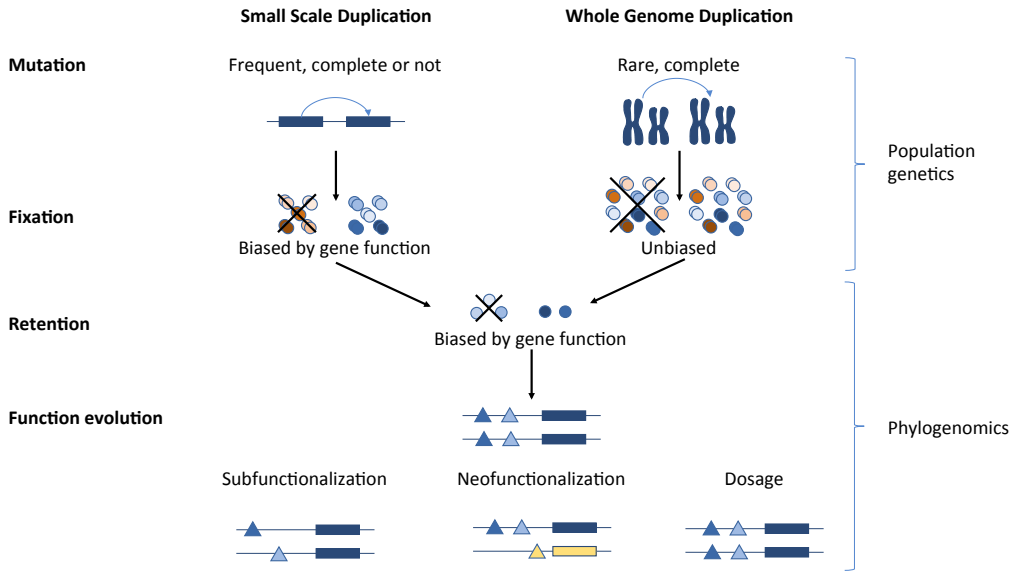


Figure 1 Dynamics of gene duplication evolution from a functional perspective. In the bottom section of the figure, the triangles represent subfunctions of each gene, for example different regulatory elements.

and Conery 2000], as common as point mutation), it has the largest impact on overall phylogenomics. The function of genes affects their duplication patterns. Functional biases can be at the mutation level (higher probability of duplicating shorter genes, or genes expressed in the germinal line), as well as fixation and retention (Figure 1). Some functional categories tend to duplicate and be lost from genomes (i.e., turn-over) much more. Other functional categories are very conservative, and are mostly found as 1-to-1 orthologs between species. Some of the same functional categories which evolve rapidly at the sequence level also have a large turn-over of gene copy number (Heger and Ponting, 2007; Ponting, 2008), notably immune defence and host evasion, and reproduction. These functions thus evolve rapidly both by amino acid substitutions and by duplication and loss of genes, allowing rapid adaptation, typically within arms-race contexts. Another functional class with abundant turn-over is metabolism genes (Demuth and Hahn, 2009), whereas these genes tend to evolve conservatively in protein sequence. Variation in copy number of metabolism genes can either contribute to the functional diversity of metabolic pathways, or to changes in dosage of metabolism proteins. Whatever the patterns of duplication, some functions seem more resistant to gene loss (Albalat and Cañestro, 2016), probably due to low dispensability of the specific function of genes in those categories. Observed patterns of gene duplication are in great part due to variations in the selection pressure that drives paralog retention or loss after the duplication event itself. From this point of view, there are important differences between whole genome duplications and small-scale duplications. All genes are duplicated in a genome duplication, and there are no issues of stoichiometry nor of missing regulatory regions for some duplicate copies. Thus the impact of gene function on retention is not biased by other processes. Studies have found long term retention of 10-20% of duplicate genes after whole genome duplication (Wolfe, 2001; Jaillon et al., 2004; Nakatani et al., 2007; Putnam et al., 2008). There is strong evidence that this loss of duplicates is non-random,

and thus enriches genomes in specific classes of genes (Davis and Petrov, 2004; Brunet et al., 2006; Roux and Robinson-Rechavi, 2008; Makino et al., 2009; Gout et al., 2010; Makino and McLysaght, 2012). In vertebrates, for example, this biased retention seems largely driven by selection against detrimental mutations of genes. This leads to a pattern of retention of genes whose variants have a higher chance of being toxic (see selection against protein misfolding above), such as those involved in diseases (Singh et al., 2014) and of genes highly expressed in the nervous system (Roux et al., 2017). While there are general trends in gene turn-over for broad categories, many specific gene family expansions or losses are lineage-specific (Lespinet et al., 2002). There are biases in gene "duplicability" which affect the small-scale duplications, which lead to such expansions, and unlike for whole genome duplication, all steps can be biased, from the duplication mutation itself to fixation, and to retention. As an example of mutation bias, there are more retrogenes from genes expressed in testis in mammals (Kaessmann et al., 2009). Fixation bias appears to go in the opposite direction for small-scale duplicate genes than for genome duplication, with genes under strong purifying selection being eliminated before fixation as paralogs (Rice and McLysaght, 2017; Roux et al., 2017). While these mechanisms are mostly due to the varying strength of purifying selection, gene family expansions of some functional categories appear to be good candidates for adaptation. For example, olfactory receptors have repeatedly expanded in lineages such as fishes, mammals, or ants (Hussain et al., 2009; Niimura et al., 2014; McKenzie and Kronauer, 2018). Gene function affects every step of the evolutionary dynamics of duplication, and ignoring the biases in generation, fixation, and retention of paralogs can lead to wrong inferences (Davis and Petrov, 2004; Studer and Robinson-Rechavi, 2009). This is a more general lesson: to study the evolution of gene function we should always control for the ways in which function can impact evolution upstream of the changes we want to study.

3 How does gene function evolve?

In addition to the impact of function on gene evolution, the function of genes itself evolves. This is in principle the most interesting aspect of the phylogenomics of function. Yet it is poorly known because this is where the difficulties in defining gene function are the most disturbing. The impact of function on gene evolution is evident through large differences between broad categories. Low granularity of functional classification is sufficient to show that immune system genes evolve under stronger positive selection, or that genes expressed in the nervous system are more often kept in several copies after genome duplication. But the evolution of gene function very rarely consists in shifts between these broad categories. Indeed, the success of gene and protein domain annotation by homology (Jiang et al., 2016) testifies to the rarity of radical shifts in function during gene evolution. Such shifts do occur, most dramatically illustrated by crystallins in tetrapod eyes (reviewed in Graur, 2016). For example in rabbits cystallin λ is a paralog of a dehydrogenase, and in frogs crystallin ρ is a paralog of a reductase. Sometimes the same protein carries both an enzymatic function and the crystallin function, known as "moonlighting proteins" (Jeffery, 2018), for example crystallin ϵ in crocodiles and ducks which is also a lactate dehydrogenase. Such cases remain rare as far as we know. Transcription factors remain transcription factors, but change subtly their specificity, affinity, or timing of expression. Membrane receptors remain receptors, but evolve different co-factors, or shift affinity for different ligands. Thus the study of the evolution of gene function is limited by our capacity to determine function of homologous genes both accurately and in an unbiased manner.

3.1 Evolution of gene expression

Gene expression patterns have consistently been a key feature used to characterize the evolution of function. Expression can be measured easily in diverse species, it is immediately comparable between genes that are otherwise very different (unlike, e.g., comparing the activity of a transcription factor and of an enzyme), and it lends itself well to modelling. With modern techniques it also lends itself well to large-scale studies, such as RNA-seq, including in non-model organisms. A notable example is the original model of sub-functionalization by Duplication-Degeneration-Complementation (DDC), which was derived from small-scale observations of gene expression in fish and mammalian development (Figure 1; section Function evolution of Force et al., 1999). While it is clear that gene function can change in evolution without change in expression pattern, a change in expression pattern between homologs can be interpreted as indicating that at least some aspect of the function has changed. In the DDC sub-functionalization model applied to expression patterns, paralogs evolve from an ancestral gene which has several domains of expression, and by losing different domains of expression in each paralog, end up recapitulating between them the ancestral pattern which neither covers entirely alone. These domains of expression can be anatomical domains (tissues, organs, cell types), timing of expression (e.g., over development), or any other aspect of expression (e.g., reaction to extrinsic signals, or sex bias). Thus for example after duplication of a gene expressed in the pectoral appendage bud and in the hindbrain in fish embryos, one paralog might conserve expression in the pectoral appendage bud, and the other in the hindbrain (this is the eng1a/b example used in Force et al., 1999). There have been many attempts to test this model, and while results have been mixed for the specific DDC model, they show that expression patterns, combined or not with information on expression levels, can be successfully used to study at least some aspects of gene function. For example, comparisons of expression patterns of genes in teleost fish after genome duplication to non-duplicate gar outgroup orthologs provided support for sub-functionalization, with typical patterns of each paralog expressed in different tissues, and the non-duplicated ortholog expressed in both (Braasch et al., 2016). The same study showed quantitative subfunctionalization, with the expression levels of two paralogs recapitulating the level of non-duplicated genes. Conversely, a study of expression of genes duplicated in the salmonid genome duplication found a dominant pattern of neo-functionalization, with one conserved paralog and one diverged: the former expressed in the same pattern as the non-duplicated ortholog, the latter expressed in different organs (Lien et al., 2016). A re-analysis of both studies indicates support for asymmetric evolution, but is not conclusive on sub- vs. neo-functionalization (Sandve et al., 2018).

3.2 The Ortholog Conjecture and the difficulty of assessing function evolution

Phylogenomics comparisons of function in the absence of duplication have been complicated, because the problems discussed in the first section of this chapter complicate defining a null expectation. Conservation of function can be measured in some cases (e.g. of expression among mammals in Brawand et al. 2011; Piasecka et al. 2012a), but distinguishing functional change from errors in the data and analysis is extremely difficult. A case study, which nicely illustrates the difficulties of studying gene function evolution at a phylogenomic scale, is the question of the "ortholog conjecture". The ortholog conjecture is the hypothesis that orthologous genes have mostly conserved function, or that their function diverges very slowly during evolution, whereas paralogous genes have mostly different functions, or that their

function diverges very rapidly during evolution (see Figure 2). While it was a foundational hypothesis of phylogenomics (Eisen, 1998), it has only started being tested systematically (and named) in the last 10 years (Studer and Robinson-Rechavi, 2009; Nehrt et al., 2011). The ortholog conjecture has been surprisingly difficult to confirm or infirm robustly, using diverse datasets and definitions of gene function.

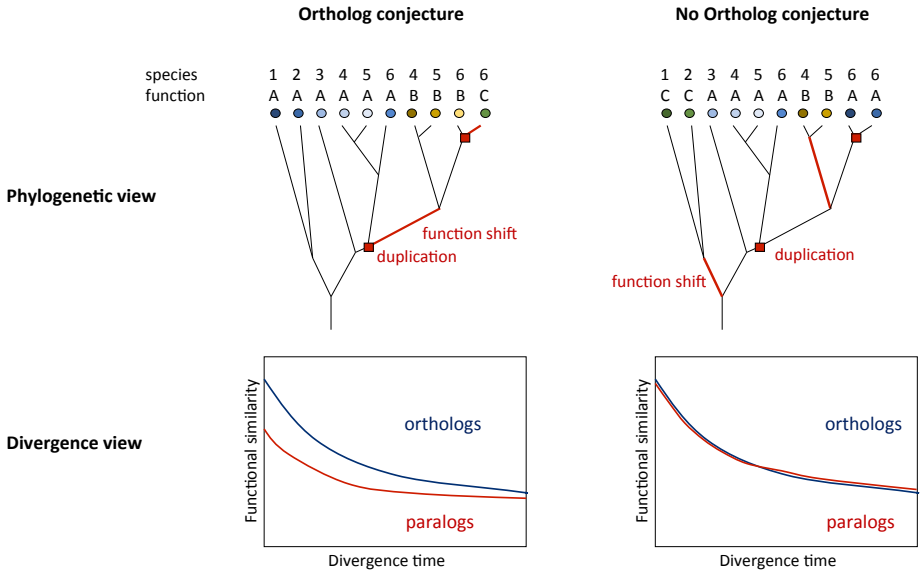


Figure 2 Schematic expectations of function evolution between orthologs and paralogs. Left, expectations under the ortholog conjecture, right, expectations if this conjecture is not supported (under a naive null of random functional changes during gene evolution). Phylogenetic view: gene tree with gene duplications indicated by red squares and functional shifts by red branches; the coloured circles are homologous genes, with the colour according to similarity of function; above, species identity (notice that following duplication, some species are represented several times in the tree) and functional classification as might be captured e.g. by the Gene Ontology. Notice that paralogs within one species might have different functions even if the ortholog conjecture is wrong, e.g. the paralogs in species 4 and 5. Divergence view: expectation of functional divergence between pairs of orthologs and of paralogs; in all cases, functional similarity is expected to decrease with evolutionary time, but paralogs are expected to diverge more and faster than orthologs under the ortholog conjecture.

Two of the first studies on the ortholog conjecture used the Gene Ontology to define functional divergence in proportion to the difference in GO annotations between genes (Nehrt et al., 2011; Altenhoff et al., 2012). Both studies took into account the ontology graph, i.e. that a hydrolase is necessarily also an enzyme, but obtained opposing results. The second study showed that paralogs in the same species tend to be studied by the same research groups, leading to similar experiments and annotations, whereas orthologs tend to be studied by different groups, leading to different experiments and annotations (see Table 1). This biases GO comparisons towards apparently more similar functional annotations between paralogs, whereas correcting for it shows more similar functional annotations between orthologs, although the effect is small (Altenhoff et al., 2012). In an unusual move, the leaders of the GO consortium published a short paper explaining why GO annotations could not be used to study evolutionary patterns of function (Thomas et al., 2012). Finally, the evolution

of GO annotations over time makes any evolutionary interpretation very difficult (Chen and Zhang, 2012). Most subsequent studies of the ortholog conjecture have focused on gene expression, for the same reasons as in other studies of gene function and evolution. Using correlations of expression levels within and between species, different studies again reached different conclusions depending on methods. Microarray data comparison was not consistent with the ortholog conjecture (Nehrt et al., 2011), but this might be due to differences in microarrays between species (Liao and Zhang, 2006; Chen and Zhang, 2012). Comparing expression levels from RNA-seq provides support for the ortholog conjecture (Chen and Zhang, 2012; Rogozin et al., 2014), although the effect size is weak and depends on the correlation method used. To avoid these issues with comparing expression levels between species, we summarized expression across tissues by the measure of "tissue-specificity", and found that it is well conserved between orthologs, different between paralogs, and diverges with time, as expected from the ortholog conjecture, and with large effect size of the difference between orthologs and paralogs (Kryuchkova-Mostacci and Robinson-Rechavi, 2016). But a reanalysis pointed out that pairwise comparisons are biased when studying evolutionary changes. Using a phylogenetic framework on the same tissue-specificity data, the support for the ortholog conjecture disappears (Dunn et al., 2018). These conflicting results show that even for a very well defined question (do paralogs diverge more than orthologs of the same age?), it is very difficult to study rigorously the evolution of gene function on a genomic scale.

4 Conclusions

The fundamental reason that we are interested in gene evolution in phylogenomics, as opposed to the evolution of random sequences of DNA, is that they carry functions, which relate the genome to the phenotype and organismal fitness. Thus we would like both to study the evolution of genes in the context of their function, allowing us to study the evolution of functional units, and to study how the function of the genes themselves evolves. On the first aim, research in the last 20 years has provided us with a view of how purifying and adaptive selection affect functional units, but limited to a very broad definition of these units: highly expressed genes, proteins central in interaction networks, potentially toxic proteins, etc. On the second aim, this lack of precision proves to be extremely limiting, and we still know surprisingly little about how gene function evolves. The difficulties in testing the "ortholog conjecture" illustrate this: if we are unable to verify such a basic assumption of our field, it seems difficult to discover new patterns until we have further improved our data and methods. Finally, the study of molecular evolution and function is in the same boat as much of genomics, suffering from too much vagueness around the notion of function (Doolittle, 2018).

References

Albalat, R. and Cañestro, C. (2016). Evolution by gene loss. *Nature Reviews Genetics*, 17(7):379–391.

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., and Dessimoz, C. (2012). Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. *PLoS Comput Biol*, 8(5):e1002514.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M.,

and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9.

Ayadi, A., Birling, M.-C., Bottomley, J., Bussell, J., Fuchs, H., Fray, M., Gailus-Durner, V., Greenaway, S., Houghton, R., Karp, N., Leblanc, S., Lengger, C., Maier, H., Mallon, A.-M., Marschall, S., Melvin, D., Morgan, H., Pavlovic, G., Ryder, E., Skarnes, W. C., Selloum, M., Ramirez-Solis, R., Sorg, T., Teboul, L., Vasseur, L., Walling, A., Weaver, T., Wells, S., White, J. K., Bradley, A., Adams, D. J., Steel, K. P., Hrabě de Angelis, M., Brown, S. D., and Herault, Y. (2012). Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mammalian Genome*, 23(9):600–610.

Barreiro, L. B. and Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics*, 11(1):17–30.

Bedford, T. and Hartl, D. L. (2009). Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences*, 106(4):1133–1138.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Braasch, I., Gehrke, A. R., Smith, J. J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A. M., Campbell, M. S., Barrell, D., Martin, K. J., Mulley, J. F., Ravi, V., Lee, A. P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Cañestro, C., Sydes, J., Beaudry, F. E. G., Sun, Y., Hertel, J., Beam, M. J., Fasold, M., Ishiyama, M., Johnson, J., Kehr, S., Lara, M., Letaw, J. H., Litman, G. W., Litman, R. T., Mikami, M., Ota, T., Saha, N. R., Williams, L., Stadler, P. F., Wang, H., Taylor, J. S., Fontenot, Q., Ferrara, A., Searle, S. M. J., Aken, B., Yandell, M., Schneider, I., Yoder, J. A., Volff, J.-N., Meyer, A., Amemiya, C. T., Venkatesh, B., Holland, P. W. H., Guiguen, Y., Bobe, J., Shubin, N. H., Di Palma, F., Alföldi, J., Lindblad-Toh, K., and Postlethwait, J. H. (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics*, 48(4):427–437.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grutzner, F., Bergmann, S., Nielsen, R., Paabo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348.

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes. *Molecular Biology and Evolution*, 23(9):1808–1816.

Chen, X. and Zhang, J. (2012). The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data. *PLoS Comput Biol*, 8(11):e1002784.

Cooper, N., Thomas, G. H., Venditti, C., Meade, A., and Freckleton, R. P. (2016). A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society*, 118(1):64–77.

Coronado-Zamora, M., Salvador-Martínez, I., Castellano, D., Barbadilla, A., and Salazar-Ciudad, I. (2019). Adaptation and Conservation throughout the Drosophila melanogaster Life-Cycle. *Genome Biology and Evolution*, 11(5):1463–1482.

Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., and Excoffier, L. (2013). Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution*.

Daub, J. T., Moretti, S., Davydov, I. I., Excoffier, L., and Robinson-Rechavi, M. (2017).

Detection of Pathways Affected by Positive Selection in Primate Lineages Ancestral to Humans. *Molecular Biology and Evolution*, 34(6):1391–1402.

Davidson, E. H. and Erwin, D. H. (2006). Gene Regulatory Networks and the Evolution of Animal Body Plans. *Science*, 311(5762):796–800.

Davis, J. C. and Petrov, D. A. (2004). Preferential Duplication of Conserved Proteins in Eukaryotic Genomes. *PLoS Biology*, 2(3):e55.

Demuth, J. P. and Hahn, M. W. (2009). The life and death of gene families. *BioEssays*, 31(1):29–39.

Dessimoz, C. and Skunca, N. (2016). *The Gene Ontology Handbook*, volume 1446 of *Methods in Molecular Biology*. Humana Press, New York, NY.

Doolittle, W. F. (2013). Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences*.

Doolittle, W. F. (2018). We simply cannot go on being so vague about 'function'. *Genome Biology*, 19(1):223.

Doolittle, W. F., Brunet, T. D. P., Linquist, S., and Gregory, T. R. (2014). Distinguishing between "function" and "effect" in genome biology. *Genome Biology and Evolution*.

Drummond, A. D. and Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet*, 10(10):715–724.

Drummond, D. A. and Wilke, C. O. (2008). Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, 134(2):341–352.

Dunn, C. W., Luo, X., and Wu, Z. (2013). Phylogenetic Analysis of Gene Expression. *Integrative and Comparative Biology*.

Dunn, C. W., Zapata, F., Munro, C., Siebert, S., and Hejnol, A. (2018). Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proceedings of the National Academy of Sciences*, 115(3):E409–E417.

Duret, L. and Mouchiroud, D. (2000). Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Mol Biol Evol*, 17(1):68–70.

Eddy, S. R. (2013). The ENCODE project: Missteps overshadowing a success. *Current biology : CB*, 23(7):R259–R261.

Eisen, J. A. (1998). Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Research*, 8(3):163–167.

Enard, D., Cai, L., Gwennap, C., and Petrov, D. A. (2016). Viruses are a dominant driver of protein adaptation in mammals. *eLife*, 5:e12469.

Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics*, 151(4):1531–1545.

Germain, P.-L., Ratti, E., and Boem, F. (2014). Junk or functional DNA? ENCODE and the function controversy. *Biology & Philosophy*, pages 1–25.

Giglio, M., Tauber, R., Nadendla, S., Munro, J., Olley, D., Ball, S., Mitraka, E., Schriml, L. M., Gaudet, P., Hobbs, E. T., Erill, I., Siegele, D. A., Hu, J. C., Mungall, C., and Chibucos, M. C. (2018). ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic Acids Research*.

Gilad, Y. and Mizrahi-Man, O. (2015). A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research*, 4:121.

Gout, J.-F., Kahn, D., Duret, L., and Paramecium Post-Genomics, C. (2010). The Relationship among Gene Expression, the Evolution of Gene Dosage, and the Rate of Protein Evolution. *PLoS Genet*, 6(5):e1000944.

Graur, D. (2016). *Molecular and genome evolution*. Sinauer Associates.

Graur, D., Zheng, Y., and Azevedo, R. B. R. (2015). An evolutionary classification of genomic function. *Genome Biology and Evolution*.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*.

Gu, X. and Su, Z. (2007). Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proceedings of the National Academy of Sciences*, 104(8):2779–2784.

He, X. and Zhang, J. (2006). Why Do Hubs Tend to Be Essential in Protein Networks? *PLoS Genetics*, 2(6):e88.

Heger, A. and Ponting, C. P. (2007). Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res.*, page gr.6249707.

Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., Proctor, M., St. Onge, R. P., Tyers, M., Koller, D., Altman, R. B., Davis, R. W., Nislow, C., and Giaever, G. (2008). The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. *Science*, 320(5874):362–365.

Ho, L. S. T. and Ané, C. (2014). Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*, 5(11):1133–1146.

Hoegg, S. and Meyer, A. (2005). Hox clusters as models for vertebrate genome evolution. *Trends in Genetics*, 21(8):421–424.

Hrycaj, S. M. and Wellik, D. M. (2016). Hox genes and evolution. *F1000Research*, 5:859.

Hueber, S. D., Weiller, G. F., Djordjevic, M. A., and Frickey, T. (2010). Improving Hox Protein Classification across the Major Model Organisms. *PLOS ONE*, 5(5):e10820.

Hurst, L. D. and Smith, N. G. C. (1999). Do essential genes evolve slowly? *Current Biology*, 9(14):747–750.

Hussain, A., Saraiva, L. R., and Korsching, S. I. (2009). Positive Darwinian selection and the birth of an olfactory receptor clade in teleosts. *Proceedings of the National Academy of Sciences*, 106(11):4313–4318.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biemont, C., Skalli, Z., Cattolico, L., Poulain, J., de Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigo, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quetier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J., and Roest Crollius, H. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957.

Jeffery, C. J. (2018). Protein moonlighting: what is it, and why is it important? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1738):20160523.

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., Koo, D. C. E., Penfold-Brown, D., Shasha, D., Youngs, N., Bonneau, R., Lin, A., Sahraeian, S. M. E., Martelli, P. L., Profiti, G., Casadio, R., Cao, R., Zhong, Z., Cheng, J., Altenhoff, A., Skunca, N., Dessimoz, C., Dogan, T., Hakala, K., Kaewphan, S., Mehryary, F., Salakoski, T., Ginter, F., Fang, H., Smithers, B., Oates, M., Gough, J., Törönen, P., Koskinen, P., Holm, L., Chen, C.-T., Hsu, W.-L., Bryson, K., Cozzetto, D., Minneci, F., Jones, D. T., Chapman, S., BKC, D., Khan, I. K., Kihara, D., Ofer, D., Rappoport, N., Stern, A., Cibrian-Uhalte, E., Denny, P., Foulger, R. E., Hieta, R., Legge, D., Lovering, R. C., Magrane, M., Melidoni, A. N., Mutowo-Meullenet, P., Pichler, K., Shypitsyna, A., Li, B., Zakeri, P., ElShal, S., Tranchevent, L.-C., Das, S., Dawson, N. L., Lee, D., Lees, J. G., Sillitoe, I., Bhat, P., Nepusz, T., Romero, A. E., Sasidharan, R., Yang, H., Paccanaro, A., Gillis, J., Sedeño-Cortés, A. E., Pavlidis, P., Feng, S., Cejuela, J. M., Goldberg, T., Hamp, T., Richter, L., Salamov, A., Gabaldon, T., Marcet-Houben, M., Supek, F., Gong, Q., Ning, W., Zhou, Y., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Toppo, S., Ferrari, C., Giollo, M., Piovesan, D., Tosatto, S. C., del Pozo, A., Fernández, J. M., Maietta, P., Valencia, A., Tress, M. L., Benso, A., Di Carlo, S., Politano, G., Savino, A., Rehman, H. U., Re, M., Mesiti, M., Valentini, G., Bargsten, J. W., van Dijk, A. D. J., Gemovic, B., Glisic, S., Perovic, V., Veljkovic, V., Veljkovic, N., Almeida-e Silva, D. C., Vencio, R. Z. N., Sharan, M., Vogel, J., Kansakar, L., Zhang, S., Vucetic, S., Wang, Z., Sternberg, M. J. E., Wass, M. N., Huntley, R. P., Martin, M. J., O'Donovan, C., Robinson, P. N., Moreau, Y., Tramontano, A., Babbitt, P. C., Brenner, S. E., Linial, M., Orengo, C. A., Rost, B., Greene, C. S., Mooney, S. D., Friedberg, I., and Radivojac, P. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(1):184.

Kachroo, A. H., Laurent, J. M., Yellman, C. M., Meyer, A. G., Wilke, C. O., and Marcotte, E. M. (2015). Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, 348(6237):921–925.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics*, 10(1):19–31.

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A. (2008). Patterns of Positive Selection in Six Mammalian Genomes. *PLoS Genetics*, 4(8):e1000144.

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2015). Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse. *PLOS ONE*, 10(6):e0131673.

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2016). Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. *PLOS Computational Biology*, 12(12):e1005274.

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2017). A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, 18(2):205–214.

Larracuente, A. M., Sackton, T. B., Greenberg, A. J., Wong, A., Singh, N. D., Sturgill, D., Zhang, Y., Oliver, B., and Clark, A. G. (2008). Evolution of protein-coding genes in Drosophila. *Trends in Genetics*, In Press, Corrected Proof.

Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes. *Genome Research*, 12(7):1048–1059.

Liao, B.-Y. and Zhang, J. (2006). Evolutionary Conservation of Expression Profiles Between Human and Mouse Orthologous Genes. *Mol Biol Evol*, 23(3):530–540.

Liao, B.-Y. and Zhang, J. (2007). Mouse duplicate genes are as essential as singletons. *Trends in Genetics*, 23(8):378–381.

Liao, B.-Y. and Zhang, J. (2008). Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proceedings of the National Academy of Sciences*, page 0800387105.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R. A., von Schalburg, K., Rondeau, E. B., Di Genova, A., Samy, J. K. A., Olav Vik, J., Vigeland, M. D., Caler, L., Grimholt, U., Jentoft, S., Inge Våge, D., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D. R., Yorke, J. A., Nederbragt, A. J., Tooming-Klunderud, A., Jakobsen, K. S., Jiang, X., Fan, D., Hu, Y., Liberles, D. A., Vidal, R., Iturra, P., Jones, S. J. M., Jonassen, I., Maass, A., Omholt, S. W., and Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, 533(7602):200–205.

Liu, J. and Robinson-Rechavi, M. (2018). Adaptive Evolution of Animal Proteins over Development: Support for the Darwin Selection Opportunity Hypothesis of Evo-Devo. *Molecular Biology and Evolution*, 35(12):2862–2872.

Lynch, M. and Conery, J. S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, 290(5494):1151–1155.

Makino, T., Hokamp, K., and McLysaght, A. (2009). The complex relationship of gene duplication and essentiality. *Trends in Genetics*, 25(4):152–155.

Makino, T. and McLysaght, A. (2012). Positionally biased gene loss after whole genome duplication: evidence from human, yeast, and plant. *Genome Research*, 22(12):2427–2435.

McDonald, A. G. and Tipton, K. F. (2014). Fifty-five years of enzyme classification: advances and difficulties. *The FEBS Journal*, 281(2):583–592.

McGary, K. L., Park, T. J., Woods, J. O., Cha, H. J., Wallingford, J. B., and Marcotte, E. M. (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences*, 107(14):6544–6549.

McKenzie, S. K. and Kronauer, D. J. C. (2018). The genomic architecture and molecular evolution of ant odorant receptors. *Genome Research*, page gr.237123.118.

Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *PNAS*, 102(31):10930–10935.

Mungall, C., Gkoutos, G., Smith, C., Haendel, M., Lewis, S., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2.

Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J. P., Jacobsen, J. O. B., Keith, D., Laraway, B., Lewis, S. E., NguyenXuan, J., Shefchek, K., Vasilevsky, N., Yuan, Z., Washington, N., Hochheiser, H., Groza, T., Smedley, D., Robinson, P. N., and Haendel, M. A. (2017). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1):D712–D722.

Nakatani, Y., Takeda, H., Kohara, Y., and Morishita, S. (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.*, page gr.6316407.

Necsulea, A. (2020). Phylogenomics and genome annotation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.1, pages 4.1:1–4.1:26. No commercial publisher | Authors open access book.

Nehrt, N. L., Clark, W. T., Radivojac, P., and Hahn, M. W. (2011). Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS Comput Biol*, 7(6):e1002073.

Niimura, Y., Matsui, A., and Touhara, K. (2014). Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome Research*, page gr.169532.113.

Noda-Garcia, L., Liebermeister, W., and Tawfik, D. S. (2018). Metabolite–Enzyme Coevolution: From Single Enzymes to Metabolic Pathways and Networks. *Annual Review of Biochemistry*, 87(1):187–216.

Obbard, D. J., Welch, J. J., Kim, K.-W., and Jiggins, F. M. (2009). Quantifying Adaptive Evolution in the Drosophila Immune System. *PLOS Genetics*, 5(10):e1000698.

Ooi, S. L., Pan, X., Peyser, B. D., Ye, P., Meluh, P. B., Yuan, D. S., Irizarry, R. A., Bader, J. S., Spencer, F. A., and Boeke, J. D. (2006). Global synthetic-lethality analysis and yeast functional profiling. *Trends in Genetics*, 22(1):56–63.

Pal, C., Papp, B., and Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet*, 37(12):1372–1375.

Pal, C., Papp, B., and Lercher, M. J. (2006). An integrated view of protein evolution. *Nat Rev Genet*, 7(5):337–348.

Pennisi, E. (2012). ENCODE Project Writes Eulogy for Junk DNA. *Science*, 337(6099):1159–1161.

Pereira, V., Waxman, D., and Eyre-Walker, A. (2009). A Problem With the Correlation Coefficient as a Measure of Gene Expression Divergence. *Genetics*, 183(4):1597–1600.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Piasecka, B., Kutalik, Z., Roux, J., Bergmann, S., and Robinson-Rechavi, M. (2012a). Comparative modular analysis of gene expression in vertebrate organs. *BMC Genomics*, 13(1):124.

Piasecka, B., Lichocki, P., Moretti, S., Bergmann, S., and Robinson-Rechavi, M. (2013). The Hourglass and the Early Conservation Models,ÄîCo-Existing Patterns of Developmental Constraints in Vertebrates. *PLoS Genet*, 9(4):e1003476.

Piasecka, B., Robinson-Rechavi, M., and Bergmann, S. (2012b). Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. *Bioinformatics*, 28(14):1865–1872.

Ponting, C. P. (2008). The functional repertoires of metazoan genomes. *Nat Rev Genet*, 9(9):689–698.

Presser, A., Elowitz, M. B., Kellis, M., and Kishony, R. (2008). The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication. *Proceedings of the National Academy of Sciences*, page 0707293105.

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. V., Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., Blow, M. J., Arkin, A. P., and Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, 557(7706):503–509.

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.-K., Benito-Gutierrez, E., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J. J., Grigoriev, I. V., Horton, A. C., de Jong,

P. J., Jurka, J., Kapitonov, V. V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L. A., Salamov, A. A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin-I, T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L. Z., Holland, P. W. H., Satoh, N., and Rokhsar, D. S. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071.

Rice, A. M. and McLysaght, A. (2017). Dosage-sensitive genes in evolution and disease. *BMC Biology*, 15(1):78.

Robinson, P. N., Köhler, S., Oellrich, A., Sanger Mouse Genetics, P., Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., Gilissen, C., Haendel, M., and Smedley, D. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, 24(2):340–348.

Robinson-Rechavi, M., Rech de Laval, V., Bastian, F. B., Wollbrett, J., and Bgee Team, p. (2020). The expression comparison tool in bgee. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.3, pages 4.3:1–4.3:4. No commercial publisher | Authors open access book.

Rocha, E. P. C. (2006). The quest for the universals of protein evolution. *Trends in Genetics*, 22(8):412–416.

Rocha, E. P. C. and Danchin, A. (2004). An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. *Mol Biol Evol*, 21(1):108–116.

Rogozin, I. B., Managadze, D., Shabalina, S. A., and Koonin, E. V. (2014). Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biology and Evolution*.

Roux, J., Liu, J., and Robinson-Rechavi, M. (2017). Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates. *Molecular Biology and Evolution*, 34(11):2773–2791.

Roux, J. and Robinson-Rechavi, M. (2008). Developmental Constraints on Vertebrate Genome Evolution. *PLoS Genetics*, 4(12):e1000311.

Roux, J., Rosikiewicz, M., and Robinson-Rechavi, M. (2015). What to compare and how: Comparative transcriptomics for Evo-Devo. *J Exp Zool B Mol Dev Evol*.

Ruff, J. S., Saffarini, R. B., Ramoz, L. L., Morrison, L. C., Baker, S., Laverty, S. M., Tvrdik, P., and Potts, W. K. (2015). Fitness Assays Reveal Incomplete Functional Redundancy of the HoxA1 and HoxB1 Paralogs of Mice. *Genetics*, 201(2):727–736.

Salvador-Martínez, I., Coronado-Zamora, M., Castellano, D., Barbadilla, A., and Salazar-Ciudad, I. (2018). Mapping Selection within Drosophila melanogaster Embryo's Anatomy. *Molecular Biology and Evolution*, 35(1):66–79.

Sandve, S. R., Rohlfs, R. V., and Hvidsten, T. R. (2018). Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature Genetics*, 50(7):908–909.

Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., and Friedberg, I. (2013). Biases in the Experimental Annotations of Protein Function and Their Effect on our Understanding of Protein Function Space. *PLoS Comput Biol*, 9(5):e1003063.

Schrempf, D. and Szöllosi, G. (2020). The sources of phylogenetic conflicts. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.1, pages 3.1:1–3.1:23. No commercial publisher | Authors open access book.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Singh, P., Affeldt, S., Cascone, I., Selimoglu, R., Camonis, J., and Isambert, H. (2012). On the Expansion of "Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates. *Cell Reports*, 2(5):1387–1398.

Singh, P. P., Affeldt, S., Malaguti, G., and Isambert, H. (2014). Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication. *PLoS Comput Biol*, 10(7):e1003754.

Sinha, S., Eisenhaber, B., Jensen, L. J., Kalbuaji, B., and Eisenhaber, F. (2018). Darkness in the Human Gene and Protein Function Space: Widely Modest or Absent Illumination by the Life Science Literature and the Trend for Fewer Protein Function Discoveries Since 2000. *PROTEOMICS*, 18(21-22):1800093.

Slodkowicz, G. and Goldman, N. (2019). Integrated evolutionary and structural analysis reveals xenobiotics and pathogens as the major drivers of mammalian adaptation. *bioRxiv*, page 762690.

Stern, D. B. and Crandall, K. A. (2018). The Evolution of Gene Expression Underlying Vision Loss in Cave Animals. *Molecular Biology and Evolution*, 35(8):2005–2014.

Studer, R. A., Penel, S., Duret, L., and Robinson-Rechavi, M. (2008). Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res.*, 18(9):1393–1402.

Studer, R. A. and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, 25:210–216.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Thomas, P. D. (2017). The Gene Ontology and the Meaning of Biological Function. In Dessimoz, C. and Škunca, N., editors, *The Gene Ontology Handbook*, Methods in Molecular Biology, pages 15–24. Springer New York, New York, NY.

Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., Blake, J. A., and on behalf of the Gene Ontology, C. (2012). On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS Comput Biol*, 8(2):e1002386.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449(7158):54–61.

Warnefors, M. and Kaessmann, H. (2013). Evolution of the Correlation between Expression Divergence and Protein Divergence in Mammals. *Genome Biology and Evolution*, 5(7):1324–1335.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet*, 2(5):333–341.

Chapter 4.3 The Expression Comparison Tool in Bgee

Marc Robinson-Rechavi, Valentine Rech de Laval, Frédéric B. Bastian, Julien Wollbrett, Bgee Team
Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland
SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
marc.robinson-rechavi@unil.ch

Abstract

We present Expression Comparison, a tool to compare expression patterns between species. It uses curated annotations of homology between anatomical structures, such as organs or tissues. Expression calls are based on the curated transcriptome data integrated within the Bgee database. Gene homology can be of any type, from user input. The results are presented according to conservation of pattern, as well as rank of expression per species. Expression Comparison is freely available on the Bgee website: `https://bgee.org`.

How to cite: Marc Robinson-Rechavi, Valentine Rech de Laval, Frédéric B. Bastian, Julien Wollbrett, Bgee Team (2020). The Expression Comparison Tool in Bgee. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 4.3, pp. 4.3:1–4.3:4. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Funding Bgee is funded by the SIB Swiss Institute of Bioinformatics.

1 Introduction

One of the most challenging aspects of phylogenomics is comparing function between homologous genes (see Chapters 4.1 and 4.2 [Necsulea 2020; Robinson-Rechavi 2020]). While gene expression is probably the aspect of function which is the most amenable to comparison between genes and between species, there has not been any tool to automatically provide such comparisons. This stands in contrast to the situation for sequences, where many databases exist which provide not only orthologous and paralogous genes, but also their sequences, multiple sequence alignments, and trees (Vilella et al., 2009; Scornavacca et al., 2019), as well as genomic regions and comparative synteny (Nguyen et al., 2018). We present here the first version of a new tool, Expression Comparison (Table 1), which leverages the Bgee database to provide such a service.

Bgee is a database of gene expression (Bastian et al., 2008, 2020) which provides manually curated healthy wild-type data for a variety of animal species, annotated to the Uberon ontology of anatomy (Haendel et al., 2014) and to standard ontologies of development and aging. Annotations also capture sex and strain or population when possible. Expression data is integrated from RNA-seq, microarrays, in situ hybridization, and ESTs. Calls of presence and absence of gene expression are made for each gene – condition combination, where a condition is a combination of anatomical structure (e.g., organ, tissue, cell type), developmental stage, sex and strain. These calls integrate all the data types together. The importance of expression of each anatomical structure in the expression of a gene is also integrated over data types, by a weighted mean of expression ranks. Thus Bgee provides a global view of "normal" gene expression in a comparable way between species, despite

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 4.3; pp. 4.3:1–4.3:4

A book completely handled by researchers.
No publisher has been paid.

differences in anatomy and data availability. These integrated calls and ranks provide a standard source of information to compare expression between genes.

To compare gene expression between species, we need to define comparable conditions. As for genes, where we seek to compare orthologs or paralogs, we need to define homologous anatomical structures to compare. In Bgee, these are manually annotated from the literature, and provided as a supplementary annotation to the Uberon ontology (Table 1). Importantly, they are integrated into the database and can be automatically queried.

Resource	Url
Expression Comparison	https://bgee.org/?page=expression_comparison
Anatomical homology annotations	https://github.com/BgeeDB/anatomical-similarity-annotations

Table 1 Resources cited in main text.

2 Using Expression Comparison

The Expression Comparison is a webserver tool which queries the Bgee database for gene expression presence or absence, rank, and anatomical homology. The user should insert into the query form a list of gene identifiers. At time of writing, it is up to the user to define this list, e.g. based on another source of information for orthology. It is possible to compare expression patterns of non homologous genes. At present, only Ensembl identifiers are accepted.

From this gene list, the tool will query Bgee and retrieve all homologous anatomical structures which have expression in at least one of the genes of the list, and have homology between the species represented by the list. Thus, a gene list with orthologs in human and zebrafish will only retrieve expression in anatomical structures which have defined homology between mammals and teleost fishes, e.g. not in placenta. The user is then provided with a list of such structures ordered by their presumed relevance; the user can re-order by clicking on column headers. A subset of results for the brain-specific gene SRRM4 (Serine/arginine repetitive matrix protein 4, required for neural cell differentiation) is presented in Table 2. The score goes from +1 for perfect conservation of expression presence, to -1 for perfect conservation of expression absence, with 0 indicating no conservation. Notice that not all 13 orthologs used in the comparison of Table 2 are always present, even for a score of 1, because there is less data in some species than others, thus some orthologs are neither called absent nor present in, e.g., cerebellar cortex. Anatomical structures are ranked by default by this score, then by genes with presence of expression, and finally by “Minimum rank”. The latter is the lowest rank of any of the compared genes in the homologous anatomical structure; a lower rank indicates a higher importance of expression. Thus the top anatomical structures reported have high consistency, have expression for many of the genes compared, and have high expression levels for at least some of these genes. Indeed, for SRRM4, the top structures are the brain and sub-parts of the brain.

User can re-order the table online. They can unfold an anatomical structure to see all the genes and the species according to their presence or absence of expression, or lack of data. The anatomical structures are linked to their description in Uberon, for users who would not know what is, e.g., Ammon’s horn. The genes are linked to their Bgee gene page, which provides detailed and species-specific expression information. The species names are linked to their Bgee species page, which provides all the data for the species in downloadable files.

Anatomical entities	Score	Minimum rank	Anatomical entity IDs	Gene count with presence of expression
brain	1	5.86E+03	UBERON:0000955	13
central nervous system	1	1.64E+04	UBERON:0001017	13
multi-cellular organism	1	1.69E+04	UBERON:0000468	13
forebrain	1	5.29E+03	UBERON:0001890	10
telencephalon	1	1.54E+04	UBERON:0001893	10
cerebellum	1	3.52E+03	UBERON:0002037	9
male reproductive system	1	2.31E+04	UBERON:0000079	6
female reproductive system	1	2.05E+04	UBERON:0000474	4
cerebellar cortex	1	3.21E+03	UBERON:0002129	3
Ammon's horn	1	1.14E+04	UBERON:0001954	3

Table 2 Top result of Expression Comparison on SRRM4 orthologs (subset of table generated from Bgee 14.1)

The table itself can be easily downloaded as TSV or copied to clipboard, for further use in, e.g., R or MS Excel.

3 Conclusion and perspectives

The Expression Comparison tool is the first available tool to automatically compare gene expression between genes taking into account curated information on anatomical homology. It leverages the data integration, transcriptome annotation, and anatomical homology annotation, in Bgee. It is already proving to be one of the more popular pages of the Bgee website, which shows that there was an unmet need for expression comparison. Future development will include automated recovery of orthologs and in-paralogs, to allow a fully automated expression comparison starting from one gene, similar to what is available in phylogenomic databases for sequences.

References

Bastian, F., Comte, A., Echchiki, A., Escoriza, A., Gharib, W., Gonzales-Porta, M., Jarosz, Y., Laurenczy, B., Mendes de Farias, T., Moret, P., Moretti, S., Niknejad, A., Parmentier, G., Person, E., Rech De Laval, V., Roelli, P., Rosikiewicz, M., Roux, J., Sanjeev, K., Seppey, M., Wollbrett, J., and Robinson-Rechavi, M. (2020). The bgee database: curated reference gene expression data and analytics tools. In preparation.

Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., and Robinson-Rechavi, M. (2008). Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In *Data Integration in the Life Sciences*, volume 5109, pages 124–131.

Haendel, M., Balhoff, J., Bastian, F., Blackburn, D., Blake, J., Bradford, Y., Comte, A., Dahdul, W., Dececchi, T., Druzinsky, R., Hayamizu, T., Ibrahim, N., Lewis, S., Mabee, P., Niknejad, A., Robinson-Rechavi, M., Sereno, P., and Mungall, C. (2014). Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of Biomedical Semantics*, 5(1):21.

Necsulea, A. (2020). Phylogenomics and genome annotation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.1, pages 4.1:1–4.1:26. No commercial publisher | Authors open access book.

Nguyen, N. T. T., Vincens, P., Roest Crollius, H., and Louis, A. (2018). Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. *Nucleic Acids Research*, 46(D1):D816–D822.

4.3:4 REFERENCES

Robinson-Rechavi, M. (2020). Molecular evolution and gene function. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.2, pages 4.2:1–4.2:20. No commercial publisher | Authors open access book.

Scornavacca, C., Belkhir, K., Lopez, J., Dernat, R., Delsuc, F., Douzery, E. J., and Ranwez, V. (2019). Orthomam v10: scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Molecular Biology and Evolution*, 36(4):861–862.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335.

Chapter 4.4 Substitution Rate Analysis and Molecular Evolution

Lindell Bromham
Macroevolution & Macroecology, Division of Ecology & Evolution
Research School of Biology Australian National University
Canberra, ACT, 0200 Australia
lndell.bromham@anu.edu.au

Abstract

The study of the tempo and mode of molecular evolution has played a key role in evolutionary biology, both as a stimulant for theoretical enrichment and as the foundation of useful analytical tools. When protein and DNA sequences were first produced, the surprising constancy of rates of change brought molecular evolution into conflict with mainstream evolutionary biology, but also stimulated the formation of new theoretical understanding of the processes of genetic change, including the recognition of the role of neutral mutations and genetic drift in genomic evolution. As more data were collected, it became clear that there were systematic differences in the substitution rate between species, which prompted further elaboration of ideas such as the generation time effect and the nearly neutral theory. Comparing substitution rates between species continues to provide a window on fundamental evolutionary processes. However, investigating patterns of substitution rates requires attention to potential complicating factors such as the phylogenetic non-independence of rates estimates and the time-dependence of measurement error. This chapter compares different analytical approaches to study the tempo and mode of molecular evolution, and considers the way a richer biological understanding of the causes of variation in substitution rate might inform our attempts to use molecular data to uncover evolutionary history.

How to cite: Lindell Bromham (2020). Substitution Rate Analysis and Molecular Evolution. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 4.4, pp. 4.4:1–4.4:21. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Substitution rates and the shape of evolutionary theory

Evolutionary genetics was founded on the patterns of inheritance of phenotypically measurable differences, and their change in frequency in populations over time. Rates of change were measured in terms of shifts in the mean trait values over time (e.g. Haldane, 1949). Mutation rates were estimated from careful detection of visible differences in members of wild populations or through laboratory crosses (e.g. Dobzhansky and Wright, 1941). While many of the leaders of the neo-Darwinian synthesis were keen to incorporate molecular data into their view of evolution, they expected it to join the party on their terms, adhering to the hard-won principle that natural selection was the composer of the molecular message, and that the genotype was servant to the phenotype (Simpson, 1964). Change in the genes and proteins, it was assumed, would reflect the changes wrought on the phenotype by selection, and would, therefore, match the phenotype in tempo and mode of evolution, varying over time as organisms responded to change in environment and selective regime (Aronson, 2002; Dietrich, 1994; Stoltzfus, 2017). Some evolutionary biologists even objected to the very notion of "molecular evolution", on the grounds that evolution as a process of phenotypic

Phylogenetics in the genomic era.
Editors: Céline Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 4.4; pp. 4.4:1–4.4:21

A book completely handled by researchers.
No publisher has been paid.

change was reflected in molecular variation, not driven by molecular change itself (Anfinsen, 1965).

Many of the pioneers of the field of molecular evolution emphasized an essentially Darwinian approach to understanding evolution, with change at the molecular level affecting short term processes of individual development as well as connecting to patterns change at longer timescales: "A general description of the evolutionary process is applicable to all levels of complexity, including the chemical level" (Dayhoff and Eck, 1969). Yet, unlike the neo-Darwinian synthesis, the new molecular view did not endow explanatory privilege on the individual level of biological organisation (e.g. Zuckerkandl and Pauling, 1965). Furthermore, it was recognized that change at the molecular level might follow different patterns to phenotypic change. In particular, the potential for neutral evolution was recognized from the beginning of the development of molecular evolution, as researchers acknowledged that change in some proteins, or parts of proteins, may be less impacted by selection than others (e.g. Anfinsen, 1959; Buettner-Janusch and Hill, 1965a). Yet these early workers were not able to make direct connections between the variation generation by mutation and the processes of evolutionary change by substitution within populations, contributing to the divergence of lineages.

In the 1960s, three new molecular techniques finally allowed scientists to peer beneath the phenotypic skin to the genotypic variation within, allowing comparison of genetic variants both within populations and between lineages. And what they saw sent shockwaves through biology. These early studies of molecular rates –from DNA hybridization, protein electrophoresis and amino acid sequences– revealed that the genome was moving out of step with the phenotype. Protein electrophoresis allowed, for the first time, some semblance of random sampling of genetic diversity within populations, measuring variability of many different proteins chosen more or less arbitrarily, revealing that a surprisingly large proportion of loci varied between individuals (Harris, 1966; Hubby and Lewontin, 1966; Lewontin and Hubby, 1966). The amount of variation at the molecular level was far higher than had been predicted from theoretical and empirical studies of rate of change at the phenotypic level (Charlesworth et al., 2016). Furthermore, DNA hybridization experiments, which used the disassociation rates of DNA from different species to indicate overall genome similarity between lineages, showed that the genome evolved continuously, and even faster than proteins. These experiments also suggested that a substantial part of the genome was made up not of unique gene sequences, each with a specific function determined by its sequence, but of vast numbers of repeats of the same short sequences (Britten and Kohne, 1968). The connection of this "repetitious DNA", if any, to the phenotype was unknown.

But the most controversial observation to come out of these early days of molecular evolutionary biology arose from the comparison of protein sequences across species. As sequences accumulated, it became possible not only to reconstruct the history of change of these molecules over evolutionary time, but also to estimate rates of change (Figure 1). Molecular change seemed to accumulate at a relatively steady rate (Doolittle and Blomback, 1964; Zuckerkandl and Pauling, 1965). This observation of constant rates was immediately put to practical use. Given an average rate of change based on fossil evidence, genetic distance between species - estimated from protein sequence comparisons, immunological distance or DNA hybridization - could be used to infer the age of their last common ancestor (Doolittle and Blomback, 1964; Zuckerkandl and Pauling, 1965; Margoliash, 1963; Wilson and Sarich, 1969).

The surprising observation of that amino acid sequences seemed to change at a roughly constant rates led to fundamental theory change, because it was used to support an argu-

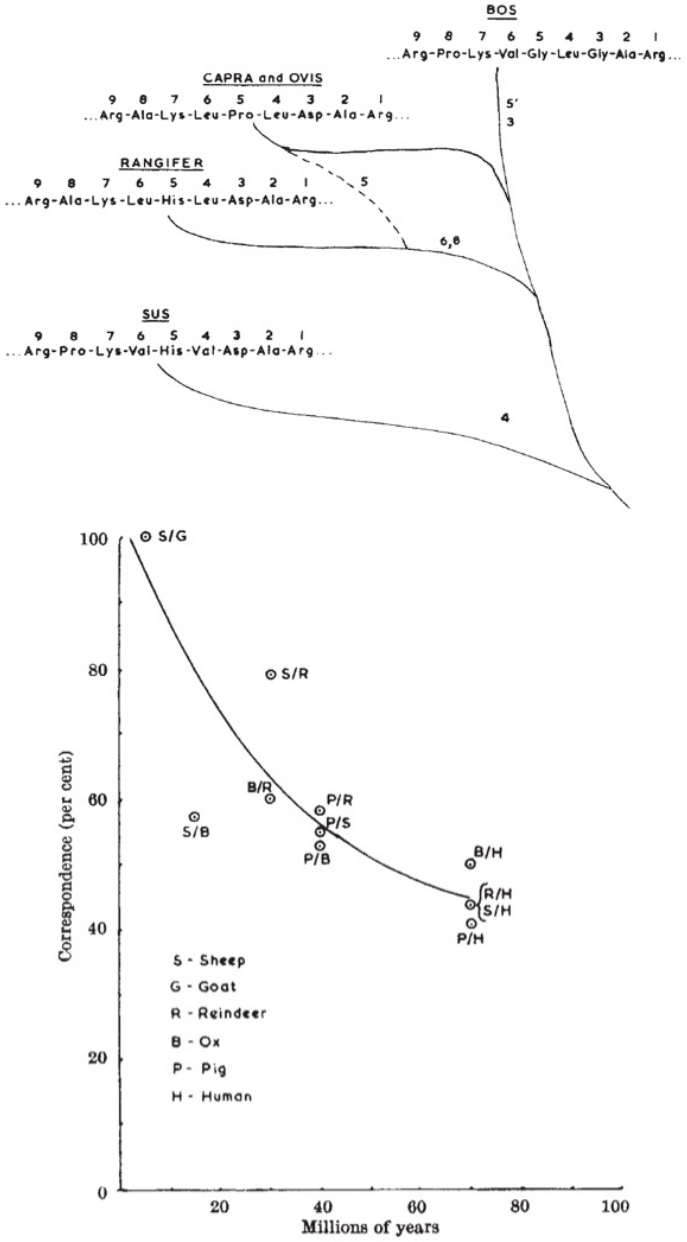


Figure 1 The advent of protein sequencing led to the first analyses of substitution rates, kicking off the controversies about molecular dating analyses that continue to this day. Reprinted by permission from Springer Nature "Amino-Acid Sequence Investigations of Fibrinopeptides from Various Mammals: Evolutionary Implications" Russell F. Doolittle, Birger Blomback Nature 1964 202(4928):147-152. Rightslink licence: 4410641268537.

ment that a substantial fraction of changes at the molecular level were neutral, and therefore not influenced by selection but only by random sampling (Kimura, 1968; King and Jukes, 1969). The possibility of neutral mutations had been recognised since the beginnings of evolutionary biology (Darwin, 1859), but had been largely rejected by whole-organism biologists (e.g. Simpson, 1964). The evolution of characters by drift was generally regarded

as of little practical impact in evolution (e.g. Fisher and Ford, 1950), or at very least as generally unproven (e.g. Cain, 1951). But those who were moving into the wild uncharted territories of comparative protein sequence analysis recognized that some changes to amino acid sequences might have no significant functional impact on the resulting protein, and might make no contribution to phenotype (Jukes 1966; Buettner-Janusch and Hill 1965b; Chapter 4.2 [Robinson-Rechavi 2020]). Such changes would not be under the influence of natural selection.

Constant rates of protein change formed one of the pillars on which the neutral theory was built (Kimura, 1968). If many mutations have little or no effect on relative fitness, then they will not be governed by selection. Their fate will be determined by chance events. Since each neutral mutation has an equal chance of drifting to fixation, their overall rate of substitution is governed by the rate at which they are generated. So Kimura (1968) proposed that the neutral substitution rate should be determined only by the neutral mutation rate.

Ironically, given the key role the molecular clock played in launching neutral theory, neutrality is neither necessary nor sufficient to explain constant rates. In fact, the apparently clock-like nature of molecular change had been debated in terms of selection for many years (e.g. Simpson, 1964), and many people working in the field were content to consider both selective and neutral explanations for constancy of rates (Zuckerkandl and Pauling, 1965). A steady rate of change could occur under selection if mutation regularly supplies variants of slight selective advantage which then undergo substitution by selection, accumulating at a roughly constant rate when considered over long time periods. Conversely, neutral evolution need not lead to constant rates. The core conclusion of the neutral theory, that the neutral substitution rate is determined by the mutation rate, leads directly to the prediction that rates of genome evolution will vary with differences in the mutation rate. It is also important to note that early molecular clock studies did not assume that the rate of change was invariant, but that any variation was random, and that the long term average rate did not differ substantially between different lineages (Margoliash, 1963). But, nonetheless, these examples show how important consideration of substitution rates has been in the debate about the causes of genomic evolution, both in the early days and continuing to the present day (e.g. Fay and Wu, 2001; Gossmann et al., 2012; Kern and Hahn, 2018; Lynch et al., 2016; Nei et al., 2010; Zhang and Yang, 2015).

In fact, it soon became apparent that rates of molecular evolution showed far more complex patterns. DNA hybridization studies revealed different rates of genomic change in different species, consistent with the prediction that species with faster generation times would generate more mutations per unit time (Laird et al., 1969; Ohta, 1972). The perceived lack of a generation time effect in protein sequence change was interpreted as a result of the interaction of several influences on rates of molecular evolution, both at the level of the mutation rate (smaller species have faster generations so generate more copy errors per year) and the substitution rate (smaller species have larger populations which have less fixation of nearly neutral changes, Ohta 1972, 1973). We now recognise a tangle of different forces that influence both mutation rate and substitution rate, which all come together to shape rates of molecular evolution, at both the DNA and protein level (Bromham, 2011).

Even in the phylogenomic era of ginormous databases, it is worth taking the time to read the earliest papers on the analysis of substitution rates, back when the challenge was to derive big theoretical conclusions from very small amounts of data (Lewontin, 1974). The foundations of the field of molecular evolution were built at a time there were few available protein sequences, each one of which had been painstakingly acquired by skilful and persistent lab work. As a consequence, a feature of this early work is the degree of

biochemical knowledge and attention to detail. Each residue that differed between species was interrogated in terms of structure and function, reactivity and charge, and interpreted in light of the principle that natural selection operates on the working properties of a three-dimensional molecule not a linear sequence of amino acids or nucleotides (Dickerson, 1971).

As the number of protein sequences grew, the first comparative databases were established. Notably, Margaret Dayhoff laid the foundations for modern phylogenomics, by bringing together biochemistry, database construction, computational tools and evolutionary principles. Her "Atlas of Protein Sequence and Structure" (Dayhoff, 1965) was the forerunner of the giant electronic databases such as GenBank. Not surprisingly, given the effort taken to generate the data, some scientists were a little possessive of their data, so Dayhoff and her collaborators had to persuade people to contribute their hard won sequences[1] (Strassman, 2012). Dayhoff also pioneered bioinformatic analysis, using computational models to examine patterns of molecular evolution (Eck and Dayhoff, 1966), constructing the first phylogeny generated through computational analysis of molecular sequences, using empirically derived frequencies to calibrate transition probabilities (Dayhoff and Eck, 1966). This work formalised the view of the sequence as a document of evolutionary history (Zuckerkandl and Pauling, 1965).

We now have so much sequence data that we are awash with information. As sequencing vast amounts of DNA becomes routine, the emphasis has shifted to large-scale computation. In only a few decades, the major challenge in molecular evolutionary biology has shifted from the problem of generating sequences and deriving evolutionary history and processes from limited data, to the problems of analysing and making sense of too much data. And so the emphasis has shifted from biochemistry to computing. As a result, we have stepped away from the sequence as representing a real molecule and are more inclined to view the sequence as a string of information. But to read the traces of evolutionary history and mechanism from the comparison of DNA, RNA or protein sequences, we need to know something of the processes that generated those traces. To do so, we need to appreciate that the sequences we analyse are a simplified representation of intricate biomolecular devices operating within living organisms, subject to a complex interacting web of biological processes and evolutionary forces. We need to remind ourselves that the string is the representation, not the reality.

2 Comparing substitution rates

Studying substitution rate is much trickier than it first appears. It would seem to be straightforward to compare sequences to come up with an estimate of the number of changes that have happened over evolutionary time from the branch lengths of a molecular phylogeny. But branch lengths reflect the amount of genetic change that has occurred, the rate at which change occurs, and the time period elapsed. None of these things is easy to measure, and often two or more of the quantities are imperfectly known, making the solution to the problem non-identifiable. If we only know only one out of the three qualities – genetic distance, time and rate – there is an infinite set of possible branch length solutions for any observed sequence data (Bromham, 2019).

For many messy problems in biology, we expect the more data we get, the more ability we

[1] As an aside, even as the gene databases expanded and went online in 1990s, many lab-based scientists who generated sequence data were somewhat reluctant to share their DNA sequences with "data parasites" who specialised in comparative analysis of sequences that other people had produced.

will have to detect signal over noise. But this is not necessarily the case with characterizing substitution rates (Bromham et al., 2017; dos Reis and Yang, 2013; Zhu et al., 2015). In fact, as the amount of data increases and the uncertainty on parameter estimates decreases, it may result in increasing confidence in the wrong answer. For example, an artefact such as long branch attraction, which can cause lineages with a rapid substitution rate to cluster together on the tree, will not necessarily be overcome by using phylogenomic datasets consisting of thousands of genes (e.g. Boussau et al., 2014; Lin et al., 2014). Given that there are many features of organismal biology that will affect the whole genome, we do not necessarily expect rate variation to simply contribute noise to substitution rate estimates, but also systematic bias. If each gene is subject to the same bias, increasing the number of loci can increase precision, but may not increase accuracy, potentially converging on the incorrect estimate (Kubatko and Degnan 2007; Kumar et al. 2012; Philippe et al. 2011; Chapter 2.1 [Simion et al. 2020]). An additional challenge with phylogenomic datasets is the possibility that loci sampled from across the genome may contain different historical narratives. Incongruence between loci may influence estimates of substitution rate, a problem that is likely to increase as more loci are analysed (Mendes and Hahn 2016; Chapters 3.3 and 3.4 [Rannala et al. 2020; Bryant and Hahn 2020]).

The problem is compounded by the evolutionary lability of rates. Substitution rates are shaped by species life history, and therefore can vary between even closely related species. To cite a few examples, rates of molecular evolution vary between mammal species according to their size, generation time, fecundity and longevity (Welch et al., 2008); closely related rockfish species can have different substitution rates if they differ in longevity (Hua et al., 2015); taller plants have slower rates of substitution (Lanfear et al., 2013); flight loss in insects leads to increased substitution rates (Mitterboeck and Adamowicz, 2013); and parasitic plants have faster rates of molecular evolution than their free-living relatives (Bromham et al., 2013). Given the large number of factors that can influence substitution rates, many of which can vary between close relatives, we expect the rate of molecular evolution to evolve as species evolve (Bromham, 2011).

Currently, there are two common approaches to dealing with evolving rates of molecular evolution when estimating substitution rates along a phylogeny for a set of sequences. One is to draw a rate for each branch independently from a convenient distribution, and choose the set of branch rates that maximizes the fit to the data, given a particular model and assumptions (generally referred to as an uncorrelated model, e.g. Drummond et al. 2006). The other is to fit an evolutionary model of rate-change to the data, allowing rates to step up and down at phylogenetic nodes or change continuously along the branches of the phylogeny (an autocorrelated rates model, e.g. Thorne et al. (1998)). All of these models are stochastic in nature and biologically arbitrary (Bromham et al., 2017). They allow rates to vary but are not informed by any special understanding of why or how they do so. There is nothing wrong with this, as long as these stochastic models can reliably capture real patterns of rate variation. But a problem arises when different rate models suggest different solutions, and we have little or no a priori information to help us decide which solution is correct (e.g. Duchêne et al., 2014; Foster et al., 2016; Lepage et al., 2007). There is some evidence that our ability to accurately infer branch length rates (i.e. distance, rates and times) using these stochastic models declines as the level of rate variation across lineages increases (Duchêne et al., 2017). In any case, the substantial variation in rate estimates generated using different methods, models and assumptions tells us that we are not yet able to precisely infer rates with the tools currently available to us. It would be helpful to have a means of studying rate variation independently of variable-rate molecular dating methods ("relaxed clocks"),

so that we can use the knowledge of patterns gained to test the validity of the relaxed clock solutions.

Estimates made independently of the relaxed clock methods may provide something of a reality-check for the phylogenetic rate estimates. Genomic analysis can provide a means of making direct estimates of rates of genome change across generations, for example by tracking genome sequence chance in microbes from lab assays (e.g. Bradwell et al., 2013), from serially sampled viruses (e.g. Duffy et al., 2008), dated ancient DNA sequences (e.g. Tong et al., 2018), or in well-studied pedigrees (e.g. Thomas et al., 2018). This direct approach to rate estimation is useful for setting empirically determined bounds on likely mutation rate values, and has been used to seek correlates of variation in rate of molecular evolution (e.g. Thomas et al., 2018). But it has its limitations. Firstly, it is applicable to only a small subset of taxa, though advances in sequencing will put pedigree analysis within reach for an increasing range of species. Secondly, mutation rates estimated in the lab or from pedigrees sometimes seem to have little direct correspondence to the values estimated from phylogenetic studies (Moorjani et al., 2016; Obbard et al., 2012), which suggests that per-generation mutation rates do not necessarily reflect long term substitution rates, even for supposedly neutral substitutions (Ho et al., 2011). Thirdly, it is important to recognise that the rates estimated from related species are likely to be more similar to each other than to randomly chosen species, due to the heritability of factors that influence mutation rate evolution (Lanfear et al., 2010). This complicates the search for consistent patterns in rate variation, because rates from different lineages cannot be treated as independent observations in a statistical analysis. So if rate estimates from each species are plotted against some other feature, such as body size or average temperature, it is not appropriate to conduct a statistical test of the association between rates and traits without correcting for covariation due to relatedness, as treating the observed rates as independent observations does not satisfy the assumption of any general statistical test such as correlation analysis (Figure 2).

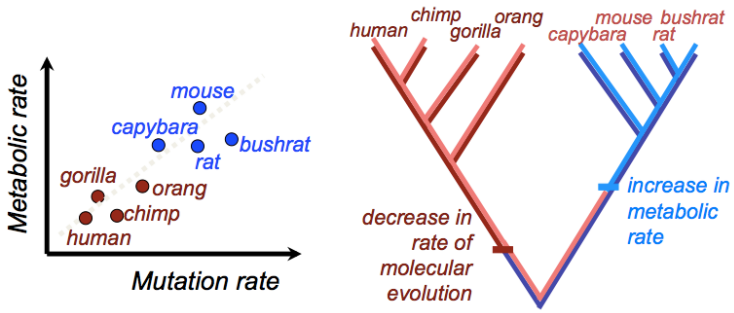


Figure 2 Why independent contrasts are necessary for the study of correlates of substitution rates. A toy example showing that if rates change along phylogenies, they can appear to be correlated with species traits that also vary between clades. In this case, because primates have undergone a slowdown in rates, rates will be correlated with anything that differs consistently between primates and rodents – for example having nails instead of claws. Reproduced from Trends in Ecology and Evolution 25, 2010 R. Lanfear, J. J. Welch, L. Bromham "Watching the Clock: studying variation in rate of molecular evolution between species" pages 495-503 with permission from Elsevier.

2.1 Phylogenetic non-independence of substitution rates

The problem of the non-independence of substitution rates due to shared descent is pertinent when we use substitution rates to answer questions about the driving forces in evolution. For example, the relationship between genome size and rate of genome change has been used to support a hypothesis that genetic drift is a major factor shaping genome evolution in organisms with small effective population sizes (Lynch and Conery, 2003; Lynch, 2010). One of the pieces of evidence provided in support of this hypothesis was a linear relationship between genome size and $N_e\mu$, a composite parameter (effective population size and mutation rate) estimated from genetic variability within a species at "silent sites". But the species-specific estimates compared in such analyses cannot be considered statistically independent observations of the influence of genome size on molecular evolution, because genome size shows phylogenetic signal in at least some groups, meaning that close relatives are more likely to have similar values than they are to randomly chosen species (e.g. Grotkopp et al., 2004; Sessegolo et al., 2016; Waltari and Edwards, 2002). Since mutation rate is influenced by species traits, it too should show phylogenetic inertia. Because species traits that could influence mutation rates, such as population size and genome size, will be more similar between relatives, this generates the potential for spurious correlations between genome size, species traits, and mutation rates (Bromham et al., 2015). A reanalysis using Phylogenetic Least Squares (PGLS) regression indicated that the significant association between effective population size, substitution rates and genome size disappears under correction for phylogenetic nonindependence (Figure 3). This does not invalidate the hypothesis, but suggests that more evidence is needed to give it empirical support, at least as far as cross-species comparisons are concerned.

If a reanalysis using methods that control for phylogenetic relatedness fails to confirm the original study, it may be tempting to conclude that the loss of significance is due to the reduction in number of datapoints reducing statistical power. A better interpretation is that the original study erroneously inflated statistical power by including data points that are effectively replicates of each other, a statistical problem that has long been recognized in evolutionary studies (including by the guy who first formulated the concept of statistical correlation analyses (Galton, 1889)). Using family-averages or including taxonomic levels as a factor in the analysis does not remove the problem of phylogenetic non-independence, because lineages within a family will still show hierarchical structuring according to relatedness, as will between-family contrasts (Bromham et al., 2018).

While there are a number of established methods for dealing with phylogenetic non-independence, care must be taken in applying standard phylogenetic comparative methods to substitution rates. Methods such as PGLS make strong assumptions about the nature of the qualities being analysed and about the way those qualities evolve over time. Specifically, they require inference of states on the internal branches, which cannot be directly measured. Typically, this involves describing the internal nodes as having values consistent with their production from a common ancestral value via a random walk along the connecting branches of the phylogenetic tree. Brownian motion is a handy way to describe random walks in character states along evolutionary trees (Lartillot and Poujol, 2011). However, there are cases of traits where this mode of inference will not provide an accurate inference of ancestral states, for example where the rate of change of traits has varied among lineages, or where there has been a directional trend in values over time (Finarelli and Flynn, 2006; Oakley and Cunningham, 2000). This may be particularly problematic for adaptive radiations such as placental mammal orders, where average body size has increased in most lineages since their last common ancestor (Bromham, 2003; Lartillot and Delsuc, 2012; Phillips, 2015). Since

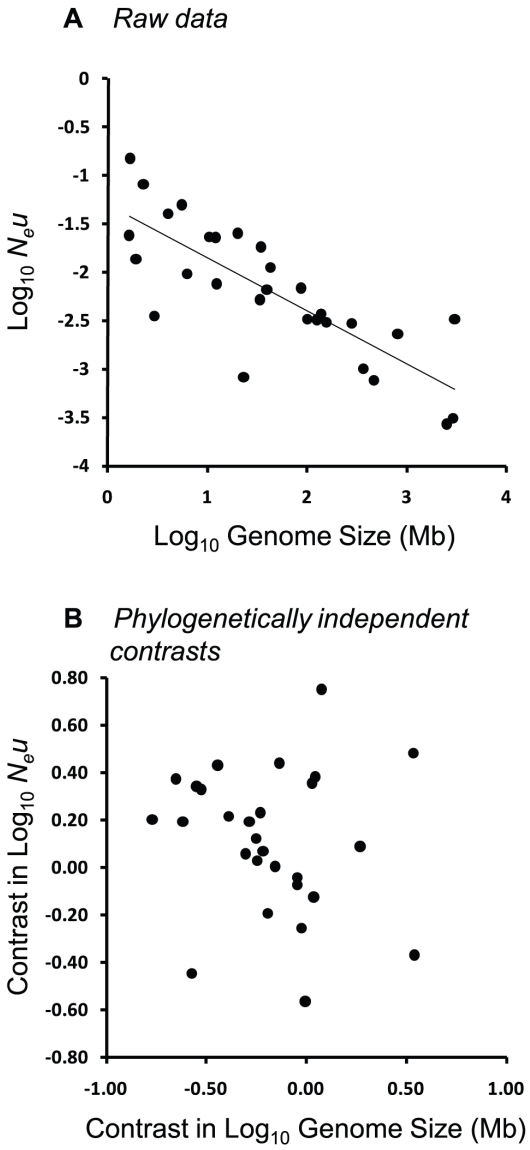


Figure 3 Controlling for phylogenetic non-independence can influence the statistical support for hypotheses about the drivers of substitution rate variation. (A) Plots of genome size against $log10N_eu$ (a composite parameter representing the effective population size and mutation rate, estimated from site variability within each species) have been used to support a causal link between genome size and rate of molecular evolution. (B) The relationship is less distinctly linear when relatedness between taxa is taken into account, and is now not statistically significant to $p < 0.05$. Reproduced under Creative Commons Attribution license (CC BY 4.0) from Whitney and Garland (2010).

substitution rates are correlated with body size in mammals, we would not expect change in rate of molecular evolution to follow a Brownian motion model for the placental mammal radiation.

PGLS and related methods are often applied to determining the patterns of molecular evolution as if substitution rates were just like other species traits, such as genome size or body mass, which can be represented as a continuous variable measured with some degree of

error. But substitution rates are something rather different. They are ultimately based on counts of changes that accrue by a stochastic process over time. This gives substitution rate measures a number of important properties that distinguish them from most other species traits, like body size or metabolic rate or niche (Welch and Waxman, 2008). One such property is that past fluctuations in rate can leave a signature in contemporary rate estimates. When we infer average ancestral states for a species trait, like body size, we typically use only the current state at the tips to derive the likely ancestral value. But, because rates reflect substitutions accruing over time, they do not really represent instantaneous measures of a trait value occurring at the tips, coincident with other species trait measurements. They represent a history of accumulation of substitutions, occurring over a protracted period of time. Because of this, a transient increase in substitution rate at some point in the past may have peppered the genome with substitutions that contribute to the estimate of substitution rate assigned to species at the tips, even after the rate has returned to the average value (Lanfear et al., 2010). For example, changes in population size over time could influence on mutation fixation rates, which might then be unrepresentative of species trait values at the tips. This is why substitution rate estimates should not be treated as if they were instantaneous measures of a species rate of genome change.

There is another property of substitution rates that sets them apart from other species traits. The accuracy of most measures of species traits is not dependent on measurements made on other species: the value of metabolic rate for a mouse is independent of whether a rat, a guinea pig or a monkey is also included in the analysis. But the estimation of the substitution rate for the mouse does depend on which other taxa are included in the analysis, as rate estimates are influenced by both the number of species included in a phylogenetic analysis and their relationship to each other. As we add in more species, we have more chance of breaking up long branches with subtending nodes, and this gives more purchase for uncovering past changes now obscured by multiple hits. More species, more nodes, more substitutions, faster rates. While the node density effect is particularly heinous for parsimony analysis, it also applies to estimates of branch length in likelihood and Bayesian methods as well. The practical upshot is that taxon sampling not only influences molecular dating analyses, but can also affect estimates of substitution rates made from phylogenies (Duchêne et al., 2015; Hugall and Lee, 2007; Linder et al., 2005; Phillips, 2015).

One approach to these problems is to simultaneously solve both rates and trait evolution for a phylogeny, then look for evidence of correlation between traits and rates over the whole tree. Whole tree analyses are increasingly being used in substitution rate analyses (e.g. Lourenco et al., 2012; Qiu et al., 2014; Santos, 2012; Wollenberg et al., 2011; Wong, 2014). While some whole-tree methods take the phylogenetic topology and branch lengths as fixed (e.g. Pagel et al., 2006), new methods jointly model rate changes and trait evolution, then assess covariation between trait and rate estimates (e.g. Lartillot and Poujol, 2011). Any use of internal edges of a phylogeny relies on being able to accurate infer past states using only the information at the tips of the tree, and this in turn relies on being able to adequately model evolutionary trajectories (typically using something like a Brownian motion model or Ornstein-Uhlenbeck process). Inference of rates changes along internal edges also requires that relative or absolute dates of divergence are known for all nodes in the phylogeny. Few phylogenies have independent dates for every node (e.g. fossil or biogeographic calibrations), so node heights must be either fixed from a molecular dating analysis (which, of course, relies on making prior assumptions about the way rates evolve over the tree), or co-estimated along with rates and traits (see Chapter 5.1 [Pett and Heath 2020]).

The flipside of using life history traits to understand evolving rates of molecular evolution is to use patterns of molecular evolution to reconstruct ancestral life histories (Wu et al., 2017). For example, models that reconstruct both life history and rates throughout the mammalian tree, using only sequences and species data derived at the tips, have led to the unexpected prediction of an ancestral placental mammal that was larger than the earliest known placental fossils, at least ten times larger than current median value for living mammals species (Jones et al., 2009), slow to mature and with a lifespan over a decade (Lartillot and Delsuc, 2012; Nabholz et al., 2013; Romiguier et al., 2013). The mammals may be a particularly challenging case study for these methods, for although the effect of life history on rates is well-studied for mammals, there is also strong directional trends in life history evolution in most placental mammalian orders (Figuet et al., 2017). For groups with a reliable fossil record, it may be possible to get more traction on evolving rates by using fossils not only to provide a prior distribution on node times, but also a prior distribution on life history traits at ancestral nodes. This would allow the estimation of ancestral rates on a phylogeny to break away from purely stochastic models and be, to some extent, ground-truthed by what we know about the biology of substitution rate variation.

2.2 Sister pairs analysis

The method of analysis that is most robust to the problems of comparative analysis of substitution rates is also the simplest[2]. If you compare the differences between two sequences that were originally copied from the same ancestral sequence, then any difference between them must have accrued since their last common ancestor. And if you have information that allows you to guess the position of that ancestor on the path of genetic change that separates them, such as an outgroup or ancestral lineage identified on a phylogeny, then you can compare the relative numbers of substitutions that have accumulated in each lineage since they split. A sister pairs approach does not produce absolute rates of change. But it does produces phylogenetically independent observations of differences in substitution rates that can be profitably used to search for correlates of rates of molecular evolution. More particularly, you can design a test where each sister pair differs in some particular trait of interest, such as life history, niche or behaviour, and you can ask whether the lineages with the greater value of the trait tend to have faster or slower rates than their sisters (Lanfear et al., 2010).

The sister pairs method has a number of advantages. Unlike PGLS and PIC, a sister pairs approach does not require a fully resolved dated phylogeny, because any information on relatedness (e.g. taxonomic information) can be used to choose non-overlapping pairs (pairs that are each others' closest relatives, with respect to any members of any other pairs in your analysis [Bromham et al. 2018]). No calibrations are required, because rates are anchored by the last common ancestor, so each member of the pair has had the same amount of time to accumulate changes. Sister pairs analyses make minimal assumptions about the model of evolution that produced the data (so, for example, they should work even when traits violate a Brownian motion model of change). Choosing a single species to represent each sister lineage removes the possibility of node density, but also forgoes the increased precision of rate estimates that comes from denser taxon sampling. Having a balanced number of taxa per sister clade should improve rate estimates, but cannot guarantee to avoid node density

[2] Which, ironically, is something of a disadvantage, as it can be hard to publish simple analyses when more complex methods are available – a kind of reverse Ockham's razor.

entirely if the distribution of speciation events or rate changes is uneven (Bromham et al., 2015; Lanfear et al., 2010). Similarly, choosing only a single locus will avoid artefacts due to gene tree discordance (Mendes and Hahn, 2016), but at the expense of including fewer informative sites.

Sister pairs analyses will solve some of the special problems of comparative analysis of rates, but not all of them. One pervasive challenge is that error in substitution rates is time-dependent, so that accurate inference of rates is tricky at both the "shallow end" and "deep end" of the evolutionary scale (e.g. van Tuinen and Torres, 2015). Systematic patterns of error in rates over time can impact on the assumptions of standard statistical tests, making correlation analyses unreliable. Accurate estimate of substitution rates from recently diverged sequences is tricky as the variance around such estimates is large due to the stochastic accrual of sequence changes. It may be tempting to dismiss poor estimates of rate due to few observable substitutions as inconsequential noise that should be overwhelmed by more robust rate estimates. But for a comparative analysis this need not necessarily be true. Welch and Waxman (2008) show how including poorly informative contrasts at the shallow end of divergence can reduce the power of comparative tests, and they recommend using simple diagnostic tests to remove these troublesome contrasts from analyses. While deleting data points can lead to a deep sense of loss, associated with the feeling that one is "throwing away data", it is preferable to being misled due to the inclusion of poor quality datapoints in an analysis, and it could lead to an ability to detect a pattern that was previously marred by the shallow datapoints (Welch and Waxman, 2008).

However, the Welch & Waxman test requires some estimate of comparison depth so that variance can be plotted against time for all contrasts. For most phylogenies, time depth comes from molecular dates, which introduces a worrying circularity for the study of the correlates of substitution rate variation. An alternative approach does not require divergence dates yet allows inclusion of shallow contrasts, by modelling the accumulation of substitutions as a Poisson process (Hua et al., 2015). The power of such comparative tests depends not only on the amount of data, but also the absolute substitution rate and also the rate of change in related species characteristics. Increasing the number of loci analysed in phylogenomic studies will help to determine the substitution rate, particularly for shallow contrasts. But for most studies, adding more independent comparisons will bring the greatest benefit in increasing the ability to detect meaningful patterns in the evolution and determination of substitution rates.

2.3 Phylogenomic data and substitution rate analysis

Phylogenomic data may help at the shallow end if including more sequence data provides a larger sample of substitutions. But it will not necessarily help at the deeper end, if too many changes have overwritten past changes. Multiple hits cause irreversible erasure of historical signal: when a site in a sequence changes more than once, the previous nucleotide states are overwritten. Overwritten history cannot be recovered, no matter how many saturated sites you look at (Bromham, 2019). Instead, we rely on models of the substitution process to guess how many changes we might no longer be able to observe, based on the pattern of those that we can see. Phylogenomic datasets may allow you a greater choice of markers to identify sequences or sites that are evolving slowly enough to avoid saturation at deep time depths, but this advantage might be lost if all loci are analysed together without discrimination. Of course, neither the deep end nor the shallow end are defined by absolute time, but by the combination of rate, time and number of observed changes (shaped by both the number of observed sites free to vary and the ability to estimate unseen changes using an evolutionary

model).

Thus far studies of correlates of substitution rates have been limited in their use of phylogenomic data. But there are many possible advantages of using a larger sample of genomic loci (Wilson Sayres et al., 2011). Multi-locus datasets provide the potential to decompose rates into gene specific and lineage-specific components (Rasmussen and Kellis, 2007). Large, genome-wide datasets may help estimate rates for shallower comparisons, allowing more meaningful comparisons between sister species. However, more loci do not necessarily provide more power to detect significant patterns in rates. For example, a phylogenomic study of rates in herbaceous and woody plants identified 5 independent comparisons between sister lineages (Yang et al., 2015). The large number of loci may provide a more comprehensive sample of sites to characterise rates across the genome, such that the rate difference for each comparison has greater confidence, but the power of the test to detect a correlation between growth habit and rates is determined by the number of independent comparisons (equivalent to the sample size in an experiment or observational study). To provide convincing test of a link between woodiness and rates, more sister comparisons would be needed, regardless of the amount of sequence data available.

3 Substitution rates shape our view of evolutionary history

The analysis of substitution rates has played an important part of developing and testing hypotheses of the drivers of molecular evolution, and the connection between change at the genotypic and phenotypic levels. But, curiously, the study of patterns of substitution rates has thusfar had relatively little impact on one of the fields where you would expect it to play a most important role. Modern molecular dating methods depend entirely on an ability to infer patterns of substitution rates over the tree, but currently the models they use are almost entirely biologically arbitrary. Very few molecular dating studies use any empirically-derived information about the way substitution rates evolve. That does not matter if our current models are up to the job. But the range of answers it is possible to get from molecular dating analyses, and the difference between published studies using the same sequence data but different methods, models and prior assumptions, suggests that we still have some way to go before we can trust molecular date estimates.

Placental mammals provide an interesting case study, for two reasons. Firstly, rates of molecular evolution have been intensively studied in mammals, and clear patterns have emerged that substitution rates are significantly associated with body size and other aspects of life history (Bromham et al., 1996; Galtier et al., 2009; Welch et al., 2008). Secondly, molecular dates for the mammalian radiation are as old as the concept of the molecular clock itself (e.g. Doolittle and Blomback, 1964; Margoliash, 1963; Zuckerkandl and Pauling, 1965), and have, for much of that history, been controversially out of step with the story told from fossil evidence alone (e.g. Bininda-Emonds et al., 2007; Hasegawa et al., 2003; Sarich and Wilson, 1967; Murphy et al., 2001). While newer molecular dating studies also tend to put the diversification of placentals in the Cretaceous, the gap between fossil and molecular dates is perceived to be shrinking (e.g. dos Reis et al., 2016; Goswami, 2012; Phillips, 2015; Ronquist et al., 2016).

This looks like a progress: more sophisticated methods and bigger datasets give us an answer that fits more snugly with both the paleontological record (fossil evidence of modern placental orders confined to post-Cretaceous) and our understanding of mammalian molecular evolution (smaller species have faster rates). It has become "a dating success story" (Goswami, 2012). But there is reason to pause for thought. The new molecular dates are

driven by two features of the new Bayesian molecular dating methods: variable-rate models and prior distributions on node height based on fossil evidence. If fossil calibrations are enforced as providing strong bounds on maximum ages, then the solution must infer very fast substitution rates on the early branches of the tree, in order to fit the sequence data to the fossil dates (O'Leary et al., 2013). If the bounds on ages are relaxed, to allow a distribution of possible ages informed by fossil data, then this allows lower rate estimates and older dates (dos Reis et al., 2014). Comparison of the prior and posterior distributions on node heights suggests that the calibrating information is strongly informative, and that estimated nodes rarely fall outside the joint prior, which may be shaped by the prior distributions on calibrations, rates, and tree shape (dos Reis et al., 2012).

It has also been suggested that the molecular dates for the placental radiation are systematically biased by uneven sampling of living mammal species, because larger-bodied contemporary species overestimate rates for the presumably smaller-bodied ancestral lineages (Phillips, 2015). A related size-biased effect has been proposed for molecular dates for the radiation of modern birds (Berv and Field, 2018). The case of the placental mammals illustrates how decisions made regarding data inclusion, calibration and other aspects of analysis can lead to substantial differences in the estimates of substitution rates and dates of divergence (e.g. dos Reis et al., 2014; Gatesy and Springer, 2017; Phillips, 2015; Springer et al., 2018; Wu et al., 2017). Despite growing confidence in molecular dating methods, there is still plenty of disagreement on molecular dates for the placental mammal radiation. So even in this case study, where we have the best understanding of the determinants of substitution rate evolution of any taxonomic group, we still have quite a long way to go before we can be sure that our molecular date estimates are not just telling us what we wanted to hear.

What has all this got to do with phylogenomics? These are systemic problems in our analysis that will not necessarily be solved by adding more data. We cannot have faith that our molecular dates will be better the more loci we include. But phylogenomic datasets give us a fantastically useful tool for understanding the way rates evolve, across the genome, over time and between lineages. The hope is that more we know about the way the historic record is written in the genome, the better we will get at reading it.

Acknowledgements

Thanks to Matt Hahn, Rob Lanfear and Xia Hua, specifically for their helpful comments on this chapter, but more generally for many wonderfully interesting and stimulating conversations on this and other topics.

References

Anfinsen, C. B. (1959). *The molecular basis of evolution*. John Wiley and Son, New York.

Anfinsen, C. B. (1965). Evolution of proteins I: Chairman's remarks. In Bryson, V. and Vogel, H. J., editors, *Evolving genes and proteins*, page 95. Academic Press, New York.

Aronson, J. D. (2002). Molecules and monkeys: George Gaylord Simpson and the challenge of molecular evolution. *History and Philosophy of the Life Sciences*, 24(3-4):441–465.

Berv, J. S. and Field, D. J. (2018). Genomic signature of an avian lilliput effect across the K-Pg extinction. *Systematic Biology*, 67(1):1–13.

Bininda-Emonds, O., Cardillo, M., Jones, K., MacPhee, R., Beck, R., Grenyer, R., Price, S.,

Vos, R., Gittleman, J., and Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, 446:507–512.

Boussau, B., Walton, Z., Delgado, J. A., Collantes, F., Beani, L., Stewart, I. J., Cameron, S. A., Whitfield, J. B., Johnston, J. S., Holland, P. W. H., Bachtrog, D., Kathirithamby, J., and Huelsenbeck, J. P. (2014). Strepsiptera, phylogenomics and the long branch attraction problem. *PLOS ONE*, 9(10):e107709.

Bradwell, K., Combe, M., Domingo-Calap, P., and Sanjuán, R. (2013). Correlation between mutation rate and genome size in riboviruses: mutation rate of bacteriophage Qβ. *Genetics*, 195(1):243–251.

Britten, R. J. and Kohne, D. E. (1968). Repeated sequences in DNA: Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science*, 161(3841):529–540.

Bromham, L. (2003). Molecular clocks and explosive radiations. *Journal of Molecular Evolution*, 57(1):S13–S20.

Bromham, L. (2011). The genome as a life-history character: Why rate of molecular evolution varies between mammal species. *Philosophical Transactions of the Royal Society of LondonB: Biological Sciences*, 366(1577):2503–2513.

Bromham, L. (2019). Six impossible things before breakfast: Assumptions, models, and belief in molecular dating. *Trends in Ecology and Evolution*, 34(5):474–486.

Bromham, L., Cowman, P. F., and Lanfear, R. (2013). Parasitic plants have increased rates of molecular evolution across all three genomes. *BMC Evolutionary Biology*, 13:126.

Bromham, L., Duchêne, S., Hua, X., Ritchie, A., Duchêne, D., and Ho, S. (2017). Bayesian molecular dating: Opening up the black box. *Biological Reviews*, 93(2):1165–1191.

Bromham, L., Hua, X., Cardillo, M., Schneemann, H., and Greenhill, S. J. (2018). Parasites and politics: why cross-cultural studies must control for relatedness, proximity and covariation. *Royal Society Open Science*, 5(8):181100.

Bromham, L., Hua, X., Lanfear, R., and Cowman, P. (2015). Exploring the relationships between mutation rates, life history, genome size, environment and species richness in flowering plants. *American Naturalist*, 185(4):507–524.

Bromham, L., Rambaut, A., and Harvey, P. H. (1996). Determinants of rate variation in mammalian DNA sequence evolution. *Journal of Molecular Evolution*, 43(6):610–621.

Bryant, D. and Hahn, M. W. (2020). The concatenation question. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.4, pages 3.4:1–3.4:23. No commercial publisher | Authors open access book.

Buettner-Janusch, J. and Hill, R. L. (1965a). Evolution of haemoglobins in primates. In Bryson, V. and Vogel, H. J., editors, *Evolving genes and proteins*, pages 167–181. Academic Press, New York.

Buettner-Janusch, J. and Hill, R. L. (1965b). Molecules and monkeys. *Science*, 147(3660):836–842.

Cain, A. J. (1951). Non-adaptive or neutral characters in evolution. *Nature*, 168:1049.

Charlesworth, B., Charlesworth, D., Coyne, J. A., and Langley, C. H. (2016). Hubby and Lewontin on protein variation in natural populations: When molecular genetics came to the rescue of population genetics. *Genetics*, 203(4):1497–1503.

Darwin, C. (1859). *The origin of species by means of natural selection: or the preservation of favoured races in the struggle for life.* John Murray, London, first edition.

Dayhoff, M. O. (1965). *Atlas of protein sequence and structure*, volume 1. National Biomedical Research Foundation, Washington.

Dayhoff, M. O. and Eck, R. V. (1966). *Atlas of protein sequence and structure*. Biomedical Research Foundation, Washington.

Dayhoff, M. O. and Eck, R. V. (1969). Inferences from protein sequence studies. In Dayhoff, M. O., editor, *Atlas of protein sequence and structure*, volume 4. Biomedical Research Foundation, Washington.

Dickerson, R. E. (1971). The structure of cytochrome c and rates of molecular evolution. *Journal of Molecular Evolution*, 1(1):26–45.

Dietrich, M. (1994). The origins of the neutral theory of molecular evolution. *Journal of the History of Biology*, 27(1):21–59.

Dobzhansky, T. and Wright, S. (1941). Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of *Drosophila pseudoobscura*. *Genetics*, 26(1):23.

Doolittle, R. F. and Blomback, B. (1964). Amino-acid sequence investigations of fibrinopeptides from various mammals: Evolutionary implications. *Nature*, 202(4928):147–152.

dos Reis, M., Donoghue, P. C., and Yang, Z. (2014). Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. *Biology Letters*, 10(1):20131003.

dos Reis, M., Donoghue, P. C., and Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, 17(2):71–80.

dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C., and Yang, Z. (2012). Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1742):3491–3500.

dos Reis, M. and Yang, Z. (2013). The unbearable uncertainty of Bayesian divergence time estimation. *Journal of Systematics and Evolution*, 51(1):30–43.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLOS Biology*, 4(5):e88.

Duchêne, D. A., Duchêne, S., and Ho, S. Y. W. (2015). Tree imbalance causes a bias in phylogenetic estimation of evolutionary timescales using heterochronous sequences. *Molecular Ecology Resources*, 15(4):785–794.

Duchêne, D. A., Hua, X., and Bromham, L. (2017). Phylogenetic estimates of diversification rate are affected by molecular rate variation. *Journal of Evolutionary Biology*, 30(10):1884–1897.

Duchêne, S., Lanfear, R., and Ho, S. Y. W. (2014). The impact of calibration and clock-model choice on molecular estimates of divergence times. *Molecular Phylogenetics and Evolution*, 78:277–289.

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*, 9(4):267–276.

Eck, R. V. and Dayhoff, M. O. (1966). Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science*, 152(3720):363–366.

Fay, J. C. and Wu, C.-I. (2001). The neutral theory in the genomic era. *Current Opinion in Genetics and Development*, 11(6):642–646.

Figuet, E., Ballenghien, M., Lartillot, N., and Galtier, N. (2017). Reconstruction of body mass evolution in the Cetartiodactyla and mammals using phylogenomic data. *bioRxiv*, page 139147.

Finarelli, J. A. and Flynn, J. J. (2006). Ancestral state reconstruction of body size in the Caniformia (Carnivora, Mammalia): The effects of incorporating data from the fossil record. *Systematic Biology*, 55(2):301–313.

Fisher, R. A. and Ford, E. B. (1950). The "Sewall Wright"effect. *Heredity*, 4:117–19.

Foster, C. S. P., Sauquet, H., Van der Merwe, M., McPherson, H., Rossetto, M., and Ho, S. Y. W. (2016). Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Systematic Biology*, 66(3):338–351.

Galtier, N., Blier, P. U., and Nabholz, B. (2009). Inverse relationship between longevity and evolutionary rate of mitochondrial proteins in mammals and birds. *Mitochondrion*, 9(1):51–57.

Galton, F. (1889). Comment on 'On a method of investigating the development of institutions; applied to laws of marriage and descent' by E. B. Tylor. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 18:245–272.

Gatesy, J. and Springer, M. S. (2017). Phylogenomic red flags: Homology errors and zombie lineages in the evolutionary diversification of placental mammals. *Proceedings of the National Academy of Sciences*, page 201715318.

Gossmann, T. I., Keightley, P. D., and Eyre-Walker, A. (2012). The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biology and Evolution*, 4(5):658–667.

Goswami, A. (2012). A dating success story: genomes and fossils converge on placental mammal origins. *EvoDevo*, 3(1):18.

Grotkopp, E., Rejmánek, M., Sanderson, M. J., and Rost, T. L. (2004). Evolution of genome size in pines (Pinus) and its life-history correlates: supertree analyses. *Evolution*, 58(8):1705–1729.

Haldane, J. B. S. (1949). Suggestions as to quantitative measurement of rates of evolution. *Evolution*, 3(1):51–56.

Harris, H. (1966). C. genetics of man enzyme polymorphisms in man. *Proceedings of the Royal Society of London B: Biological Sciences*, 164(995):298–310.

Hasegawa, M., Thorne, J. L., and Kishino, H. (2003). Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes and Genetic Systems*, 78(4):267–283.

Ho, S. Y. W., Lanfear, R., Bromham, L., Phillips, M. J., Soubrier, J., Rodrigo, A. G., and Cooper, A. (2011). Time-dependent rates of molecular evolution. *Molecular Ecology*, 20(15):3087–3101.

Hua, X., Cowman, P., Warren, D., and Bromham, L. (2015). Longevity is linked to mitochondrial mutation rates in rockfish: a test using Poisson regression. *Molecular Biology and Evolution*, 32(10):2633–2645.

Hubby, J. L. and Lewontin, R. C. (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, 54(2):577–594.

Hugall, A. F. and Lee, M. S. Y. (2007). The likelihood node density effect and consequences for evolutionary studies of molecular rates. *Evolution*, 61(10):2293–2307.

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., Safi, K., Sechrest, W., Boakes, E. H., and Carbone, C. (2009). PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, 90(9):2648–2648.

Jukes, T. H. (1966). *Molecules and evolution*. Columbia Univesrity Press, New York.

Kern, A. D. and Hahn, M. W. (2018). The neutral theory in light of natural selection. *Molecular Biology and Evolution*, 35(6):1366–1371.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626.

King, J. L. and Jukes, T. H. (1969). Non-Darwinian evolution. *Science*, 164(3881):788–798.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24.

Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky P., S. L., and Tamura, K. (2012). Statistics and truth in phylogenomics. *Molecular Biology and Evolution*, 29(2):457–472.

Laird, C. D., McConaughy, B. L., and McCarthy, B. J. (1969). Rate of fixation of nucleotide substitutions in evolution. *Nature*, 224:149 – 154.

Lanfear, R., Ho, S. Y. W., Davies, T. J., Moles, A. T. Aarssen, L., Swenson, N. G., Warman, L., Zanne, A. E., and Allen, A. P. (2013). Taller plants have lower rates of molecular evolution: the rate of mitosis hypothesis. *Nature Communications*, 4(1):1879.

Lanfear, R., Welch, J. J., and Bromham, L. (2010). Watching the clock: Studying variation in rates of molecular evolution. *Trends in Ecology and Evolution*, 25(9):495–503.

Lartillot, N. and Delsuc, F. (2012). Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, 66(6):1773–1787.

Lartillot, N. and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular Biology and Evolution*, 28(1):729–744.

Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. (2007). A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution*, 24(12):2669–80.

Lewontin, R. C. (1974). *The genetic basis of evolutionary change*. Columbia University Press, New York.

Lewontin, R. C. and Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in *Drosophila pseudoobscura*. *Genetics*, 54(2):595–609.

Lin, J., Chen, G., Gu, L., Shen, Y., Zheng, M., Zheng, W., Hu, X., Zhang, X., Qiu, Y., Liu, X., and Jiang, C. (2014). Phylogenetic affinity of tree shrews to glires is attributed to fast evolution rate. *Molecular Phylogenetics and Evolution*, 71:193–200.

Linder, H. P., Hardy, C. R., and Rutschmann, F. (2005). Taxon sampling effects in molecular clock dating: an example from the African Restionaceae. *Molecular Phylogenetics and Evolution*, 35(3):569–582.

Lourenco, J. M., Glemin, S., Chiari, Y., and Galtier, N. (2012). The determinants of the molecular substitution process in turtles. *Journal of Evolutionary Biology*, 26:38–50.

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, 26(8):345–352.

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., and Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17:704.

Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *Science*, 302(5649):1401–1404.

Margoliash, E. (1963). Primary structure and the evolution of cytochrome c. *Proceedings of the National Academy of Sciences*, 50(4):672–679.

Mendes, F. K. and Hahn, M. W. (2016). Gene tree discordance causes apparent substitution rate variation. *Systematic Biology*, 65(4):711–721.

Mitterboeck, T. F. and Adamowicz, S. J. (2013). Flight loss linked to faster molecular evolution in insects. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1767):20131128.

Moorjani, P., Gao, Z., and Przeworski, M. (2016). Human germline mutation and the erratic evolutionary clock. *PLOS Biology*, 14(10):e2000744.

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., and de Jong, W. W. (2001). Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, 294(5550):2348–2351.

Nabholz, B., Uwimana, N., and Lartillot, N. (2013). Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. *Genome Biology and Evolution*, 5(7):1273–1290.

Nei, M., Suzuki, Y., and Nozawa, M. (2010). The neutral theory of molecular evolution in the genomic era. *Annual Review of Genomics and Human Genetics*, 11:265–289.

Oakley, T. H. and Cunningham, C. W. (2000). Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution*, 54(2):397–405.

Obbard, D. J., Maclennan, J., Kim, K.-W., Rambaut, A., O'Grady, P. M., and Jiggins, F. M. (2012). Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Molecular Biology and Evolution*, page mss150.

Ohta, T. (1972). Evolutionary rate of cistrons and DNA divergence. *Journal of Molecular Evolution*, 1(2):150–157.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96–98.

O'Leary, M. A., Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., Goldberg, S. L., Kraatz, B. P., Luo, Z.-X., Meng, J., Ni, X., Novacek, M. J., Perini, F. A., Randall, Z. S., Rougier, G. W., Sargis, E. J., Silcox, M. T., Simmons, N. B., Spaulding, M., Velazco, P. M., Weksler, M., Wible, J. R., and Cirranello, A. L. (2013). The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science*, 339(6120):662–667.

Pagel, M., Venditti, C., and Meade, A. (2006). Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science*, 314(5796):119–121.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLOS Biology*, 9(3):e1000602.

Phillips, M. J. (2015). Geomolecular dating and the origin of placental mammals. *Systematic Biology*, 65(3):546–557.

Qiu, F., Kitchen, A., Burleigh, J. G., and Miyamoto, M. M. (2014). Scombroid fishes provide novel insights into the trait/rate associations of molecular evolution. *Journal of Molecular Evolution*, 78(6):338–348.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Rasmussen, M. D. and Kellis, M. (2007). Accurate gene-tree reconstruction by learning gene-and species-specific substitution rates across multiple complete genomes. *Genome Research*, 17(12):1932–1942.

Robinson-Rechavi, M. (2020). Molecular evolution and gene function. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.2, pages 4.2:1–4.2:20. No commercial publisher | Authors open access book.

Romiguier, J., Ranwez, V., Douzery, E. J. P., and Galtier, N. (2013). Genomic evidence for large, long-lived ancestors to placental mammals. *Molecular Biology and Evolution*, 30(1):5–13.

Ronquist, F., Lartillot, N., and Phillips, M. J. (2016). Closing the gap between rocks and clocks using total-evidence dating. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 371(1699):20150136.

Santos, J. C. (2012). Fast molecular evolution associated with high active metabolic rates in poison frogs. *Molecular Biology and Evolution*, 29(8):2001–2018.

Sarich, V. M. and Wilson, A. C. (1967). Immunological time scale for hominid evolution. *Science*, 158(805):1200.

Sessegolo, C., Burlet, N., and Haudry, A. (2016). Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biology Letters*, 12(8):20160407.

Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

Simpson, G. G. (1964). Organisms and molecules in evolution. *Science*, 146(3651):1535–1538.

Springer, M. S., Murphy, W. J., and Roca, A. L. (2018). Appropriate fossil calibrations and tree constraints uphold the Mesozoic divergence of solenodons from other extant mammals. *Molecular Phylogenetics and Evolution*, 121:158–165.

Stoltzfus, A. (2017). Why we don't want another 'synthesis'. *Biology Direct*, 12(1):23.

Strassman, B. J. (2012). Dayhoff, M. O. In *Encyclopedia of Life Sciences*. Wiley.

Thomas, G. W. C., Wang, R. J., Puri, A., Harris, R. A., Raveendran, M., Hughes, D., Murali, S., Williams, L., Doddapaneni, H., and Muzny, D. (2018). Reproductive longevity predicts mutation rates in primates. *Current Biology*, 28(19).

Thorne, J. L., Kishino, H., and Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12):1647–1657.

Tong, K. J., Duchêne, D. A., Duchêne, S., Geoghegan, J. L., and Ho, S. Y. W. (2018). A comparison of methods for estimating substitution rates from ancient DNA sequence data. *BMC Evolutionary Biology*, 18(1):70.

van Tuinen, M. and Torres, C. R. (2015). Potential for bias and low precision in molecular divergence time estimation of the canopy of life: an example from aquatic bird families. *Frontiers in Genetics*, 6:203.

Waltari, E. and Edwards, S. V. (2002). Evolutionary dynamics of intron size, genome size, and physiological correlates in archosaurs. *The American Naturalist*, 160(5):539–552.

Welch, J. J., Bininda-Emonds, O. R. P., and Bromham, L. (2008). Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evolutionary Biology*, 8:53.

Welch, J. J. and Waxman, D. (2008). Calculating independent contrasts for the comparative study of substitution rates. *Journal of Theoretical Biology*, 251(4):667–678.

Whitney, K. D. and Garland, T., J. (2010). Did genetic drift drive increases in genome complexity? *PLOS Genetics*, 6(8):e1001080.

Wilson, A. C. and Sarich, V. M. (1969). A molecular timescale for human evolution. *Proceedings of the National Academy of Sciences*, 63(4):1088–1093.

Wilson Sayres, M. A., Venditti, C., Pagel, M., and Makova, K. D. (2011). Do variations in substitution rates and male mutation bias correlate with life-history traits? a study of 32 mammalian genomes. *Evolution*, 65(10):2800–2815.

Wollenberg, K. C., Vieites, D. R., Glaw, F., and Vences, M. (2011). Speciation in little: the role of range and body size in the diversification of Malagasy mantellid frogs. *BMC Evolutionary Biology*, 11(1):217.

Wong, A. (2014). Covariance between testes size and substitution rates in primates. *Molecular Biology and Evolution*, 31(6):1432–1436.

Wu, J., Yonezawa, T., and Kishino, H. (2017). Rates of molecular evolution suggest natural history of life history traits and a post-K-Pg nocturnal bottleneck of placentals. *Current Biology*, 27(19):3025–3033.e5.

Yang, Y., Moore, M. J., Brockington, S. F., Soltis, D. E., Wong, G. K.-S., Carpenter, E. J., Zhang, Y., Chen, L., Yan, Z., Xie, Y., Sage, R. F., Covshoff, S., Hibberd, J. M., Nelson, M. N., and Smith, S. A. (2015). Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution*, 32(8):2001–2014.

Zhang, J. and Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16:409–420.

Zhu, T., Dos Reis, M., and Yang, Z. (2015). Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. *Systematic Biology*, 64(2):267–280.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*, 97:97–166.

Chapter 4.5 Detecting Adaptation from Multi-species Protein-coding DNA Sequence Alignments

Christine Lowe
Agriculture and Agri-Food Canada
Biological Informatics Centre of Excellence, Ottawa, ON, Canada
christine.lowe@canada.ca

Nicolas Rodrigue
Carleton University
Department of Biology, Institute of Biochemistry, and School of Mathematics and Statistics, Ottawa, ON, Canada
nicolas.rodrigue@carleton.ca

Abstract

Modern methods to detecting adaptive evolution from interspecific protein-coding gene alignments rely on statistical models of sequence evolution formulated at the level of codons. By performing model comparisons, one measures the evidence for signals of adaptive substitution processes, relative to a null model that disallows any adaptive regime. In this chapter, we present the detailed form of these models of sequence evolution, and how they are applied to real data sets. The classical codon substitution models are based on evaluating the relative nonsynonymous to synonymous substitution rates, and the main focus has traditionally been placed on devising models allowing for increasingly more subtle manifestations of adaptive substitution processes. We also overview a contrasting modeling direction that has emerged in the last decade—although with roots two decades back—in which the emphasis is placed on devising a richer modeling of purifying selection. Using simulations, we expand the characterization of this latter approach, followed by a contrasting of its conclusions on real data with those of classical codon models. Finally, we discuss the numerous model violations that can lead to erroneous inferences on various tests, and potential future directions meriting attention.

How to cite: Christine Lowe and Nicolas Rodrigue (2020). Detecting Adaptation from Multi-species Protein-coding DNA Sequence Alignments. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 4.5, pp. 4.5:1–4.5:18. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Introduction

When alignments of protein-coding DNA sequences from different species became available in the second half of the 20th century, evolutionary biologists quickly sought to contrast the rate of substitutions that do not alter the encoded amino acid sequence (the *synonymous* substitutions) to those that imply an amino acid replacement (the *nonsynonymous* substitutions). Early methods were based on simple counting schemes (Miyata and Yasunaga, 1980; Perler et al., 1980; Gojobori, 1983; Li et al., 1985; Nei and Gojobori, 1986). One of their objectives was to account for the fact that not all codon states have the same potential for synonymous and nonsynonymous substitutions; for instance, a codon encoding tryptophan has no synonymous opportunity, given that it is alone in encoding this amino acid, whereas leucine is encoded by six codons, and therefore has high synonymous opportunity. These

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 4.5; pp. 4.5:1–4.5:18

A book completely handled by researchers.
No publisher has been paid.

early methods also relied on multiple pairwise sequence comparisons, for all or most of the possible pairs from the multiple sequence alignment. By the 1990s, however, a number of statistical models were proposed, working within a full phylogenetic framework (Goldman and Yang, 1994; Muse and Gaut, 1994; Halpern and Bruno, 1998). Rather than utilizing counting schemes on pairwise comparisons, the models were based on the idea of fitting parameters (e.g., by maximum likelihood) representing features of a codon substitution process running over the branches of a phylogenetic tree relating all sequences of an alignment.

In this chapter, we follow two main threads in the development of codon substitution models aimed at detecting adaption in protein-coding genes. The first consists of models based on a parameter denoted ω, which corresponds to the rate ratio of nonsynonymous (dN) and synonymous (dS) substitutions ($\omega = dN/dS$). The traditional interpretation of ω is that instances where model fitting leads to $\omega \sim 1$ correspond to (nearly) neutral evolution, whereas $\omega < 1$ indicates purifying, or negative selection, and $\omega > 1$ corresponds to an adaptive regime, or positive selection. We present the detailed form of such models, as well as the commonly used extensions allowing for variation in ω across codon sites of an alignment. The second modeling thread we present is focused on better capturing the subtleties of purifying selection, in what has come to be known as the *mutation-selection* framework. These approaches attempt to account for the heterogeneity of amino acid fitness profiles across sites, and form a null model against which to test for nonsynonymous rates greater than expected under the mutation-selection balance. We describe the mutation-selection rationale in detail, and present a simulation study to further characterize the approach. We also contrast its results on several thousands of real data sets with those of the classical codon substitution models. Finally, we discuss a number of model violations that can influence inferences under codon substitution models, and outline future research directions that merit greater attention.

2 The classic codon substitution models

Appearing back-to-back in the 1994 September issue of Molecular Biology and Evolution, the papers by Muse and Gaut (1994, "MG") and Goldman and Yang (1994, "GY") took the ideas of likelihood-based phylogenetic analysis in the nucleotide state space, and proposed to expand the state space to in-frame nucleotide triplets: rather than specifying a 4 by 4 matrix of nucleotide substitution rates, a 61 by 61 (assuming a universal genetic code) matrix of codon substitution rates is specified; the lethality of stop-codons is a built-in assumption of the model, in being disallowed in the state space. Another built-in assumption is that of a point-mutation process, where the substitution rate between codons that differ by two or three nucleotides is set to zero. However, this latter assumption is not new to the codon-level context, but inherent to the nucleotide-level context as well, since the probability of two nucleotide sites undergoing a substitution within a given time interval vanishes as the interval approaches zero (see Chapter 1.1 [Pupko and Mayrose 2020]).

2.1 MG-style models

Thanks to the point-mutation assumption, one can re-formulate a nucleotide-level model, such as the general-time-reversible (GTR) model (Lanave et al., 1984), into a nucleotide triplet state space. Let $\rho = (\rho_{lm})_{1 \leq l,m \leq 4}$ be a set of (symmetrical) nucleotide relative exchangeability parameters, with the constraint $\sum_{1 \leq l < m \leq 4} \rho_{lm} = 1$. Also let $\varphi = (\varphi_m)_{1 \leq m \leq 4}$, with $\sum_{m=1}^{4} \varphi_m = 1$, be a set of nucleotide equilibrium frequency parameters. The GTR model specifies the entries of a 4 by 4 rate matrix as $Q_{lm} = \rho_{lm}\varphi_m$. The exact same model

can be written into a 64 by 64 matrix, specifying rates from one codon i to another j as:

$$Q_{ij} = \begin{cases} \rho_{i_c j_c} \varphi_{j_c}, & \text{if } i \text{ and } j \text{ differ only at } c^{\text{th}} \text{ codon position,} \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where i_c corresponds to an index of the nucleotide at the c^{th} codon position ($c = 1$, 2, or 3) of codon i ($i_c = 1$, 2, 3, or 4, as indices for A, C, G, and T). The distinction between Equation 1 and the GTR model is only one of a game of indices, but the formulation given in Equation 1 suggests that we could further recognize different types of codon substitutions. For instance, if we suppress stop codons from the process (reducing it to a 61 by 61 rate matrix, as stated above), we could recognize the distinction between synonymous and nonsynonymous substitution rates, specifying the entries in the matrix as:

$$Q_{ij} = \begin{cases} \rho_{i_c j_c} \varphi_{j_c}, & \text{if } i \text{ and } j \text{ are synonymous} \\ \rho_{i_c j_c} \varphi_{j_c} \omega, & \text{if } i \text{ and } j \text{ are nonsynonymous} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The model given in Equation 2 resembles closely the one proposed by Muse and Gaut (1994). The most notable difference is that in their seminal paper, Muse and Gaut invoked a new multiplicative parameter for nonsynonymous and synonymous rates, whereas in Equation 2 we effectively set the synonymous rate multiplier to 1, and invoke a single parameter, ω, on nonsynonymous rates, and hence $\omega = dN/dS$. The other difference with the original model is that Equation 2 includes parameters controlling nucleotide exchangeabilities (as in a GTR nucleotide-level model), whereas Muse and Gaut originally did not, and only used the frequency parameter of the target nucleotide (as in a F81 nucleotide-level model, Felsenstein, 1981). The stationary probability of codon i, denoted π_i, which can be thought of as the proportion of time spent in codon state i when running the substitution process for a very long time, is given by:

$$\pi_i = \frac{\varphi_{i_1} \varphi_{i_2} \varphi_{i_3}}{\sum_j \varphi_{j_1} \varphi_{j_2} \varphi_{j_3}}, \quad (3)$$

where the summation in the denominator is over all 61 (non-stop) codon states. In other words, the stationarity of the model given by Equation 2 is nearly identical to what it would be under the GTR model over three nucleotide positions, but with a slight re-normalization for the absence of the stop codons from the state space.

One of the widely used extensions to this formulation, often denoted as F3x4, is to invoke three distinct nucleotide frequency vectors, $\varphi^{(1)}$, $\varphi^{(2)}$, and $\varphi^{(3)}$, for the three within-codon positions (Yang, 2006), and hence with a rate matrix given by:

$$Q_{ij} = \begin{cases} \rho_{i_c j_c} \varphi^{(c)}_{j_c}, & \text{if } i \text{ and } j \text{ are synonymous} \\ \rho_{i_c j_c} \varphi^{(c)}_{j_c} \omega, & \text{if } i \text{ and } j \text{ are nonsynonymous} \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

and stationary probability given by:

$$\pi_i = \frac{\varphi^{(1)}_{i_1} \varphi^{(2)}_{i_2} \varphi^{(3)}_{i_3}}{\sum_j \varphi^{(1)}_{j_1} \varphi^{(2)}_{j_2} \varphi^{(3)}_{j_3}}. \quad (5)$$

The idea behind F3x4 formulation—so called because it involves three sets of four-dimensional frequency vectors—is to account for the uneven frequencies observed across the three within-codon positions that result from the structure of the genetic code; for instance, if there is

a highly skewed compositional bias in the data, it is most reflected in the third codon positions, whose state typically has no impact on the encoded amino acid, whereas first and second codon positions, whose states generally dictate the amino acid, would have different nucleotide frequencies. In other words, the differences in nucleotide frequencies across the three positions are features of different selective pressures operating at the amino acid level. While it may seem artificial to have an enriched parameterization at the nucleotide-level to account for features of selection at the amino acid level, the justification is phenomenological: regardless of the mechanistic details at the amino acid level, this modeling approach attempts to capture the net effect of these mechanisms, at least partially.

Pond et al. (2010) proposed a correction to the common practice of setting the values of these parameters to the nucleotide frequencies observed at the three codon positions of the data set at hand. The parameters can also be estimated by maximum likelihood, or become part of a Bayesian inference (Rodrigue et al., 2008a).

2.2 GY-style models

Models inspired by Goldman and Yang (1994) differ from those inspired by Muse and Gaut (1994) in a few ways. Perhaps the most significant difference, however, is the fact that with GY-style models the entries in the substitution rate matrix are proportional to the frequency (or stationary probability) of the target *codon*:

$$Q_{ij} = \begin{cases} \rho_{i_c j_c} \pi_j, & \text{if } i \text{ and } j \text{ are synonymous} \\ \rho_{i_c j_c} \pi_j \omega, & \text{if } i \text{ and } j \text{ are nonsynonymous} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

This contrasts with the MG-style models given in Equation 2, which have entries proportional to the frequency of the target *nucleotide* (φ_{j_c}).

Most practitioners set the values of the 61-dimensional π-vector based on nucleotide-level specifications. For instance, it is common to work with a single vector of nucleotide frequencies, denoted F1x4, and setting $\pi_i \propto \varphi_{i_1}\varphi_{i_2}\varphi_{i_3}$. This practice, however, leads to peculiar effects in the specification of codon substitution rates, that could sometimes contradict the modeling intention. Rodrigue et al. (2008a) point out an example: suppose a context that is highly sucesptible to events leading to A or to T (in other words, a context where φ_A and φ_T are at higher values than φ_C and φ_G); all else being equal, the substitution rate from CGC to CTC (i.e., toward a high-frequency state T) will be lower than the rate from ATA to AGA (i.e., toward a low-frequency state G). Stated in reciprocally, the rate will be higher for the event that goes against the compositional bias at the nucleotide level (to a final state G), simply because of the context of the event. It is unclear if practitioners fully realize these sorts of peculiarities of GY-style F1x4 models, or if they consider them to be negligible to the inference of interest, usually ω (or its distribution).

Naturally, the F3x4 configuration, setting $\pi_i \propto \varphi^{(1)}_{i_1}\varphi^{(2)}_{i_2}\varphi^{(3)}_{i_3}$ is also utilized extensively in GY-style models. However, Huelsenbeck and Dyer (2004), as well as Rodrigue et al. (2008b), have shown that such approximations of π are often well outside the 95% credibility intervals of full Bayesian inferences of the 61-dimensional vector (a setting often denoted F61). On the other hand, the main drawback of the GY-F61 model is the confounded account of nucleotide, amino acid, and codon propensities. As discussed in a later section, the MG-style models offer opportunities to account for these propensities separately, within the mutation-selection framework.

3 Distributions of ω

Beyond issues relating to MG versus GY, or F1x4, F3x4, and F61, the main objective of codon substitution models is to characterize the ratio of nonsynonymous rates to synonymous rates, or ω, as denoted above, with a particular interest in cases where $\omega > 1$. In the previous section, we presented the classic codon substitution models in their homogeneous versions; each codon column of the alignment is considered to be the realization of a strictly identical Markov process, remaining unchanged across the branches of the phylogeny, or across the positions of the alignment. When fitting such global models to real data, one virtually never encounters cases where $\omega > 1$, simply because adaptive evolution is unlikely to be operating across all states, sites, and branches. In this section, we present the ideas behind the most widely known models of codon substitutions that account for variation in ω, with a focus on across-site heterogeneity.

3.1 Variable ω across sites

The first models to account for variation in ω values across the codon alignment were inspired from the random effects approaches of Yang (1993, 1994) to model rates across sites in nucleotide-level models. They consider each codon column of the alignment to have been produced from a model with ω drawn from a parametric statistical law (Nielsen and Yang, 1998; Yang et al., 2000). A problem remains, however, in that there is no inherent reason to choose one statistical law over another. The approach taken in the seminal works of Yang, Nielsen and collaborators was empirical: explore many different statistical laws, and perform likelihood-based model comparisons to identify the most appropriate one.

The simplest strategy to capturing an unknown distribution is to discretize it, into a finite mixture model. Suppose that we allow for K different ω parameters operating across codon alignment sites, denoted ω_1, ω_2,..., ω_K. The probability of the nth codon alignment column, denoted D_n, given the parameters of the model, denoted collectively as θ, is given as a weighted average over the K components of the mixture:

$$p(D_n \mid \theta) = \sum_{k=1}^{K} w_k p(D_n \mid \theta, \omega_k), \tag{7}$$

where $w = (w_k)_{1 \leq k \leq K}$, with $\sum_{k=1}^{K} w_k = 1$ is a set of weights associated with each component; these weights can be thought of as the prior probabilities that a particular alignment codon column D_n was generated by each of the K components.

In what they refer to as their *neutral* model, Nielsen and Yang (1998) set $K = 2$, $\omega_1 = 0$, and $\omega_2 = 1$, and infer the remaining parameters by maximum likelihood. In other words, their neutral model assumes a mixture of two codon sites, one in which the nonsynonymous substitution rate matches the synonymous substitution rate, and one in which nonsynonymous events are disallowed. Their *positive selection* model adds a third class ($K = 3$), with $\omega_3 > 1$. A likelihood ratio test can be performed between these two models, to establish if there is evidence of positive selection. Such a test is an example of the general approach to detecting adaptive evolution in the maximum likelihood framework, often followed by Empirical Bayes methods for identifying sites with high probability of having $\omega > 1$ (reviewed in Anisimova, 2012).

Capturing the unknown distribution of ω-values across sites can also be explored using continuous parametric distributions. For instance, rather than a two-component neutral model where sites either belong to a component with $\omega_1 = 0$ or $\omega_2 = 1$, one could invoke a

continuous distribution in the [0, 1] range, such as the beta distribution (Yang et al., 2000). Under such a model, a site likelihood takes the following form:

$$p(D_n \mid \theta) = \int_{\omega_n} p(\omega_n \mid \alpha, \beta) p(D_n \mid \theta, \omega_n) d\omega_n, \tag{8}$$

where $p(\omega_n \mid \alpha, \beta)$ is the density under the beta distribution (analogous to w_k in Equation 7), parameterized by α and β (which become part of the ML inference). In practice, the integral in Equation 8 is approximated through a discretization technique, reducing it to a weighted sum much like in Equation 7, but the overall approach still allows for a compact parameterization accounting for the heterogeneity across sites.

Of course, one could also envisage models built from a mixture of a continuous distribution and a discrete category $\omega_p > 1$ to allow for sites with positive selection, leading to a site likelihood with the form:

$$p(D_n \mid \theta) = w_1 \left(\int_{\omega_n} p(\omega_n \mid \alpha, \beta) p(D_n \mid \theta, \omega_n) d\omega_n \right) + w_2 p(D_n \mid \theta, \omega_p). \tag{9}$$

As before, this latter model, along with the previous one based on the beta distribution alone, can form the basis of a likelihood ratio test for the presence of sites with signatures of adaptive evolution.

Models based on mixtures of discrete and continuous distributions, or several continuous distributions, can be built in a similar fashion, which led Yang et al. (2000) to propose the suite of M-class models, with associated likelihood-ratio tests for adaptive substitution regimes. Indeed, the latter test based on comparing a model invoking a beta distribution and a additional component $\omega_p > 1$ against a model with only the beta distribution is known as the M8 versus M7 test.

These types of models have also been studied in the Bayesian context (Huelsenbeck and Dyer, 2004), within which they were also extended into non-parametric versions based on the Dirichlet process (Huelsenbeck et al., 2006; Rodrigue et al., 2008b,a).

3.2 Increasingly subtle modeling of variation in ω

Historically, the development of codon models has mainly progressed in the manner described above: the focus has been on proposing models that would allow for increasingly subtle manifestations of adaptive evolution, such as adaptive regimes operating at particular sites, and/or along particular branches (Yang and Nielsen, 2002; Yang et al., 2005; Yang and Dos Reis, 2010; Guindon et al., 2004), typically through the use of a variety of statistical devices controlling the values of ω across sites and branches. This focus may have turned attention away from a richer modeling of mutational features, as well as impeded a more general questioning of the use of ω as an appropriate means of detecting adaptation.

Indeed, the interpretation of ω values has been a point of contention (Nielsen and Yang, 2003; Seo and Kishino, 2008). Nielsen and Yang (2003) made indirect connections between this parameter and basic population genetics theory. They related the value of ω to the *scaled selection coefficient*, denoted S, which quantifies the change in fitness associated to a particular amino acid replacement, through the expression $\omega = S/(1 - e^{-S})$. This relation raises some odd scenarios. For instance, $\omega > 1$ implies that all nonsynonymous substitutions (at a given site and/or branch) have $S > 0$; an amino acid replacement from 'L' to 'I' would have a positive selection coefficient, and so would one from 'I' to 'L'. Conversely, cases where $\omega < 1$ imply that every nonsynonymous substitution decreases fitness ($S < 0$), including, say, 'D' to 'E' *and* 'E' to 'D'.

Meanwhile, another modeling rationale had emerged with an entirely different focus: devising a better representation of the pervasive underlying purifying selection operating on protein-coding DNA sequences.

4 The mutation-selection framework

In 1998, Halpern and Bruno (1998, "HB') introduced a model formulation with a more straightforward interpretation directly rooted in population genetics concepts. Their basic idea was to explicitly recognize both the initial and final states of a codon substitution, rather than simply distinguishing between synonymous and nonsynonymous events, and the identity of the final nucleotide or codon state. Yang and Nielsen (2008) offered a clear presentation of the idea of the model, introducing a fitness parameter f_i for codon i. A change from a wild-type state i to a mutant state j then implies a selection coefficient $s_{ij} = f_j - f_i$. The fixation probability associated to the mutant is given (approximated) by $2s_{ij}/(1 - e^{-2N_e s_{ij}})$, where N_e is the effective chromosomal population size. In a context involving haploids, N_e directly corresponds to the effective population size, whereas with diploids, our notation implies that N_e is twice the effective population size; in other words, the N_e we refer to here includes a ploidy-dependent multiplicative factor (see Yang and Nielsen, 2008, for details). A mutational process can be specified, for instance as a nucleotide-level GTR model, where the mutation rate from codon i to j is given by $\mu_{ij} = \rho_{i_c j_c} \varphi_{j_c}$, and the chromosomal population-level mutation rate is thus $N_e \mu_{ij}$. These two concepts are combined multiplicatively to specify the substitution rate:

$$Q_{ij} = \begin{cases} N_e \mu_{ij} \frac{2s_{ij}}{1-e^{(-2N_e s_{ij})}}, & \text{if } i \text{ and } j \text{ differ by one nucleotide,} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

By multiplying the leftmost N_e factor through the fixation probability, and replacing $2N_e s_{ij}$ with S_{ij}, the scaled selection coefficient (scaled by twice the effective chromosomal population size), the model given by Equation 10 can be re-written as:

$$Q_{ij} = \begin{cases} \mu_{ij} \frac{S_{ij}}{1-e^{-S_{ij}}}, & \text{if } i \text{ and } j \text{ differ by one nucleotide,} \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $S_{ij} = F_j - F_i$, and where $F_i = 2N_e f_i$ is the scaled fitness of codon i. One can estimate scaled fitness parameters by anchoring one of them at $F_i = 0$ (for instance), and estimating the remaining fitness parameters around this constraint; what matters is the *relative* scaled fitness. An alternative, however is to set $F_i = \ln \psi_i$, where $\psi = (\psi_i)_{1 \leq i \leq 61}$, with $\sum_{i=1}^{61} \psi_i = 1$, is a codon profile. With this mutation-selection framework, the interpretations of $S_{ij} < 0$, $S_{ij} = 0$, and $S_{ij} > 0$ as negative, neutral, and positive selection coefficients apply to specific events, as opposed to being the coefficients of a long-standing regime implied from Nielsen and Yang (2003)'s interpretation.

The stationary probability of codon i under such a model is given by

$$\pi_i = \frac{\varphi_{i_1} \varphi_{i_2} \varphi_{i_3} e^{F_i}}{\sum_j \varphi_{j_1} \varphi_{j_2} \varphi_{j_3} e^{F_j}}, \tag{12}$$

or equivalently

$$\pi_i = \frac{\varphi_{i_1} \varphi_{i_2} \varphi_{i_3} \psi_i}{\sum_j \varphi_{j_1} \varphi_{j_2} \varphi_{j_3} \psi_j}. \tag{13}$$

In Equations 12 and 13, we have defined parameters controlling nucleotide propensities, φ, and parameters controlling codon fitness, ψ. The stationary probability, π, is calculated on the basis of φ and ψ, but the presentation of the model suggests that it is φ and ψ (or F) that are being estimated. In the presentation of the model by Halpern and Bruno (1998), however, it is π and φ that are estimated, with ψ (or F) being implicit. Specifically, they construct the substitution matrix as:

$$Q_{ij} = \begin{cases} \mu_{ij} \dfrac{\ln\left(\frac{\pi_j \mu_{ji}}{\pi_i \mu_{ij}}\right)}{1-\left(\frac{\pi_i \mu_{ij}}{\pi_j \mu_{ji}}\right)}, & \text{if } i \text{ and } j \text{ differ by one nucleotide,} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

The equivalence of these approaches is made plain by noting that substituting Equation 13 into Equation 14 yields Equation 11.

4.1 HB-style models

Above, the form of the mutation-selection framework is presented as a global model. A core idea of Halpern and Bruno HB model, however, is to have a unique set of codon profiles for each site. Moreover, in practice, they reduce the model to having only amino acid profiles; this yields a model for each site n given by:

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij} \dfrac{\ln \phi_{f(j)}^{(n)} - \ln \phi_{f(i)}^{(n)}}{1-\left(\phi_{f(i)}^{(n)} / \phi_{f(j)}^{(n)}\right)}, & \text{if } i \text{ and } j \text{ are nonsyn. and differ by one nucleotide,} \\ \mu_{ij}, & \text{if } i \text{ and } j \text{ are syn. and differ by one nucleotide,} \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where $\phi^{(n)} = (\phi_a^{(n)})_{1 \le a \le 20}$ is the amino acid profile operating at site n, and $f(i)$ returns an index for the amino acid encoded by codon i. Note that with $S_{ij}^{(n)} = \ln \phi_{f(j)}^{(n)} - \ln \phi_{f(i)}^{(n)}$, we can re-write the model in manner similar to Equation 11:

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij} \dfrac{S_{ij}^{(n)}}{1-e^{-S_{ij}^{(n)}}}, & \text{if } i \text{ and } j \text{ are nonsyn. and differ by one nucleotide,} \\ \mu_{ij}, & \text{if } i \text{ and } j \text{ are syn. and differ by one nucleotide,} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

As such, the HB model is one that specifies as many codon substitution matrices as there are codon columns in the alignment, with each sharing a single nucleotide-level parameterization, but having a unique set of amino acid fitness parameters to it alone.

The adoption of the approach was hindered by its very high dimensionality, with a decade passing before such site-specific parameters were used again (Holder et al., 2008; Tamuri et al., 2012, 2014). There are also concerns with the treatment of site-specific profiles as *bona fide* parameters to be estimated by ML, given that such conditions do not conform with those of asymptotic theory of likelihood inference (Rodrigue, 2013): site-specific approaches do not have asymptotic conditions, since applying them to data sets with increasingly greater number of sites implies introducing new amino acid profiles, and thus changing the model; applying them to data sets with more sequences also changes the model, by requiring additional branch lengths, over a different tree. An alternative modeling strategy, routinely applied in most phylogenetic analyses, is the random variable approach.

4.2 Random-variable approaches for mutation-selection models

As described earlier in the context of classical codon models focused on ω, the random variable approach has also been invoked for amino acid fitness profiles, with both parametric (Rodrigue, 2013) and non-parametric (Rodrigue et al., 2010b; Rodrigue and Lartillot, 2014) methods. As before, the amino acid fitness profiles are considered random variables, integrated over a statistical law (Lartillot 2006; Chapter 1.4 [Lartillot 2020]). Along the parametric versions, the statistical laws utilized in previous studies have included a plain flat Dirichlet on the 20 amino acid states (Rodrigue and Aris-Brosou, 2011; Rodrigue, 2013), a free Dirichlet, itself with parameters controlling its center and concentration (Lartillot, 2006; Rodrigue, 2013), or finite mixture models with empirically derived values (Rodrigue and Aris-Brosou, 2011; Kazmi and Rodrigue, 2019).

Along the non-parametric versions, the Dirichlet process on amino acid fitness profiles has been utilized, implemented via both "Chinese restaurant" (Rodrigue et al., 2010b) and "stick-breaking" (Rodrigue and Lartillot, 2014) representations. Both representations utilize an auxiliary variable $z = (z_n)_{1 \leq n \leq N}$, specifying for each codon site the current allocation to one of K sets of "active"[1] amino acid fitness profiles, and the substitution model at site n is often presented as:

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij} \frac{S_{ij}^{(z_n)}}{1 - e^{-S_{ij}^{(z_n)}}}, & \text{if } i \text{ and } j \text{ are nonsyn. and differ by one nucleotide,} \\ \mu_{ij}, & \text{if } i \text{ and } j \text{ are syn. and differ by one nucleotide,} \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

where $S_{ij}^{(z_n)} = \ln \phi_{f(j)}^{(z_n)} - \ln \phi_{f(i)}^{(z_n)}$ is the scaled selection coefficient based on the amino acid profile allocated to site n, denoted $\phi^{(z_n)}$. The model given in Equation 17 is sometimes denoted MutSelDP.

In spite of being referred to as *non-parametric*, the Dirichlet process indeed involves parameters (often referred to as *hyper-parameters*), specifying a *base distribution*, analogous to a mean, and a *granularity* parameter, controlling the coarseness of the estimation of the unknown distribution. The likelihood function can be thought of as an integral over the infinite set of mixture models, conditional on these hyperparameters; the integral is effectively approximated using Monte Carlo methods. To date, relatively little work has been done to study the Dirichlet process with alternative hyperparameters; previous studies have endowed them with their own simple statistical laws (hyperpriors), and treated them as free elements of the inference. Richer hyperpriors, such as a base distribution itself consisting of a mixture of Dirichlets, should be studied in future work.

4.3 The mutation-selection framework as a null model for detecting adaptation

The basic motivation behind the mutation-selection framework set out by Halpern & Bruno is to define a better null model as a starting point to understanding features of the evolution of protein-coding genes (Rodrigue et al., 2010b). Specifically, the framework is focused on capturing purifying selection in a site-heterogeneous manner. Spielman and Wilke (2015) clearly lay out out how these models have the effect of inducing a dN/dS ratio less than 1

[1] The Monte Carlo devices invoke large sets of amino acid profiles, some of which are not actually allocated to any sites in the alignment (see Lartillot et al., 2013, for details).

at stationarity. We refer to the dN/dS ratio induced by the mutation-selection formulation as ω_0, and it is intuitively straightforward to see why it is constrained in the 0 to 1 range, by examining two extreme cases. First, suppose that, for some reason, all amino acids have nearly the same fitness; then, all values of S_{ij} would be close to 0, and thus $S_{ij}/(1-e^{-S}) \sim 1$, such that, when accounting for mutational opportunity (see Spielman and Wilke, 2015, for details), the nonsynonymous rate and the synonymous rate closely match (i.e., ω_0 is close to 1). At the other extreme, suppose that the amino acid fitness profile is strongly dominated by a single amino acid; then, at stationarity, it would be rare to be in a state other than this dominating amino acid, with all possible mutations away from it leading to very negative values of S_{ij}, and thus $S_{ij}/(1-e^{-S}) \sim 0$, such that the induced nonsynonymous rate is close to 0 (i.e., ω_0 is close to 0). Other configurations on amino acid profiles, between these two extreme scenarios, lead to ω_0 in the 0 to 1 range.

These ideas suggest an alternative approach to detecting adaptive regimes: rather than aiming to detect cases where $\omega > 1$, we could aim to detect cases where the overall dN/dS is greater than what would be expected under a pure mutation-selection formulation. One approach to this is to introduce a multiplicative parameter to nonsynonymous events, which we denote ω_* (the asterisk is used to clearly distinguish this parameter from ω), leading to the following model by Rodrigue and Lartillot (2017):

$$Q_{ij}^{(n)} = \begin{cases} \mu_{ij}\omega_* \frac{S_{ij}^{(zn)}}{1-e^{-S_{ij}^{(zn)}}}, & \text{if } i \text{ and } j \text{ are nonsyn. and differ by one nucleotide,} \\ \mu_{ij}, & \text{if } i \text{ and } j \text{ and syn. and differ by one nucleotide,} \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The model given in Equation 18 is referred to as MutSelDP-ω_*. With such a formulation, the overall $\omega = dN/dS$ of the model at stationarity can be thought of as $\omega = \omega_*\omega_0$. Or, written alternatively as, $\omega_* = \omega/\omega_0$, this new parameter can be thought of as a measure of the deviation in the overall dN/dS (ω) with respect to what is expected from the pure mutation-selection formulation (ω_0). A value of $\omega_* > 1$ signals that the overall nonsynonymous rate is greater than expected under the mutation-selection balance, indicating a potential adaptive regime. The approach is much less demanding than classical codon models requiring that the nonsynonymous rate actually surpasses the synonymous rate, and could therefore have the potential to detect manifestations of adaptation that would otherwise be overlooked. More fundamentally, the approach demonstrates a different modeling perspective: that of formulating a better null framework, in order to uncover more subtle deviations from this new null. Such ideas were also studied by Bloom (2017).

5 Simulation study

Rodrigue and Lartillot (2017) conducted a brief simulation study to highlight the general behaviour of the MutSelDP-ω_* model given in Equation 18 when encountering data generated under an adaptive regime. Without going into the details of the simulation methods (described in full in Rodrigue and Lartillot, 2017), the idea is to change the amino acid fitness parameters (used to evolve sequences) over a phylogenetic tree; at a certain *rate* (ρ in Rodrigue and Lartillot, 2017) over the branches of the phylogeny, the amino acid profiles are altered (referred to as a *Red Queen* regime in Rodrigue and Lartillot, 2017). Thus, a sequence evolving along the branches with such changes in amino acid profiles is never quite at equilibrium, since it is tracking a repeatedly changing fitness optimum. In order to achieve a state of higher fitness, the sequence will accumulate a greater number of nonsynonymous substitutions than it would have in the absence of changes in fitness over time.

Note, however, that the increased nonsynonymous rate is generally still far from surpassing the synonymous rate.

In addition to controlling the rate of the Red Queen—the rate at which amino acid profiles are changed over the phylogeny—we can control the *magnitude* (σ_{RQ} in Rodrigue and Lartillot, 2017) of the change brought about to entries in the amino acid profile being altered. We can also control the *degree* (K in Rodrigue and Lartillot, 2017) of the change in profile, which corresponds to the number of pairs of entries in a profile that are being changed. We also alter the *proportion of sites* that are evolved under a Red Queen, with the remaining sites evolved under a pure mutation-selection process. Finally, we have an absolute *mutation rate*, which acts as a multiplier in the data-generating substitution matrices, and controls the overall amount of evolutionary signal in the simulations (set at 2×10^{-4} in Rodrigue and Lartillot, 2017). The simulations originally presented by Rodrigue and Lartillot (2017) only altered the rate of the Red Queen, leaving all other parameters of the simulation set to arbitrary values. Here, we further explore how the model reacts to a range of different values for other simulation parameters, with the results displayed as box-plots of the posterior mean values of ω_* over 20 replicate simulations in Figure 1. As done by Rodrigue and Lartillot (2017), we varied the rate of the Red-Queen (Figure 1A,B), however, rather than applying the Red-Queen regime to all sites, we explore results with the proportion of sites in an adaptive regime ranging from 0 to 1. We also adjust the number of pairs altered, or degreE,From 1 to 10 (Figure 1E,F), and study different magnitudes of changes brought to each pair of profiles (Figure 1C,D). All simulation scenarios were repeated on the basis of two starting sets of amino acid fitness profiles: those obtained from running the plain MutSelDP model (without ω_*) as given in Equation 17 on the BRCA1 alignment described by Rodrigue and Lartillot (2017), or on the concatenated alignment by Lartillot and Delsuc (2012).

In examining all of the performed simulations, increasing the proportion of sites in the alignment evolving under a Red-Queen evolutionary regime led to progressively higher ω_* values (Figure 1). This is easy to understand, as ω_* is a global parameter, estimated on the basis of the joint information found in the alignment. Therefore, if only a few sites experience an adaptive process, the model will accommodate the majority of the positions. As more sites are under a Red-Queen regime, the value of ω_* increases, and if the Red-Queen is sufficiently pronounced, the probability that ω_* is greater than 1 given the data, written symbolically as $p(\omega_* > 1 \mid D)$, approaches 1. When the Red-Queen regime is applied to 10% of sites, only 8% of simulations were found to be under positive selection using the BRCA1 amino acid profiles and 6% using the nuclear concatenation-based profiles ($p(\omega_* > 1 \mid D) \geq 0.95$). When all sites are subject to the Red-Queen regime, 89% of simulations were found to be under positive selection based on the BRCA1-derived profiles and 93% based on the concatenation-derived profiles.

Simulations based on concatenation- and BRCA1-derived profiles exhibited similar overall trends. We note, however, that in simulations based on the concatenation-derived profiles without any sites evolving under a Red-Queen, ω_* appears to be slightly below 1 (Figure 1B,D,F). Under null conditions, 2% of concatenation-based simulations where above 1, similar to the 1% observed with the BRCA1-based simulations. More noteworthy, under null conditions, 10% of BRCA1-based simulations were below 1 and 18% of concatenation-based simulations ($p(\omega_* < 1 \mid D) \geq 0.95$). The tendency of these simulations to lead to ω_* values below 1 supports the conclusions of Spielman and Wilke (2015) that the current MutSelDP model overestimates ω_0. However, null-generated data leading to ω_* values less than 1 suggests that detecting adaptive evolution using $\omega_* > 1$ will be conservative.

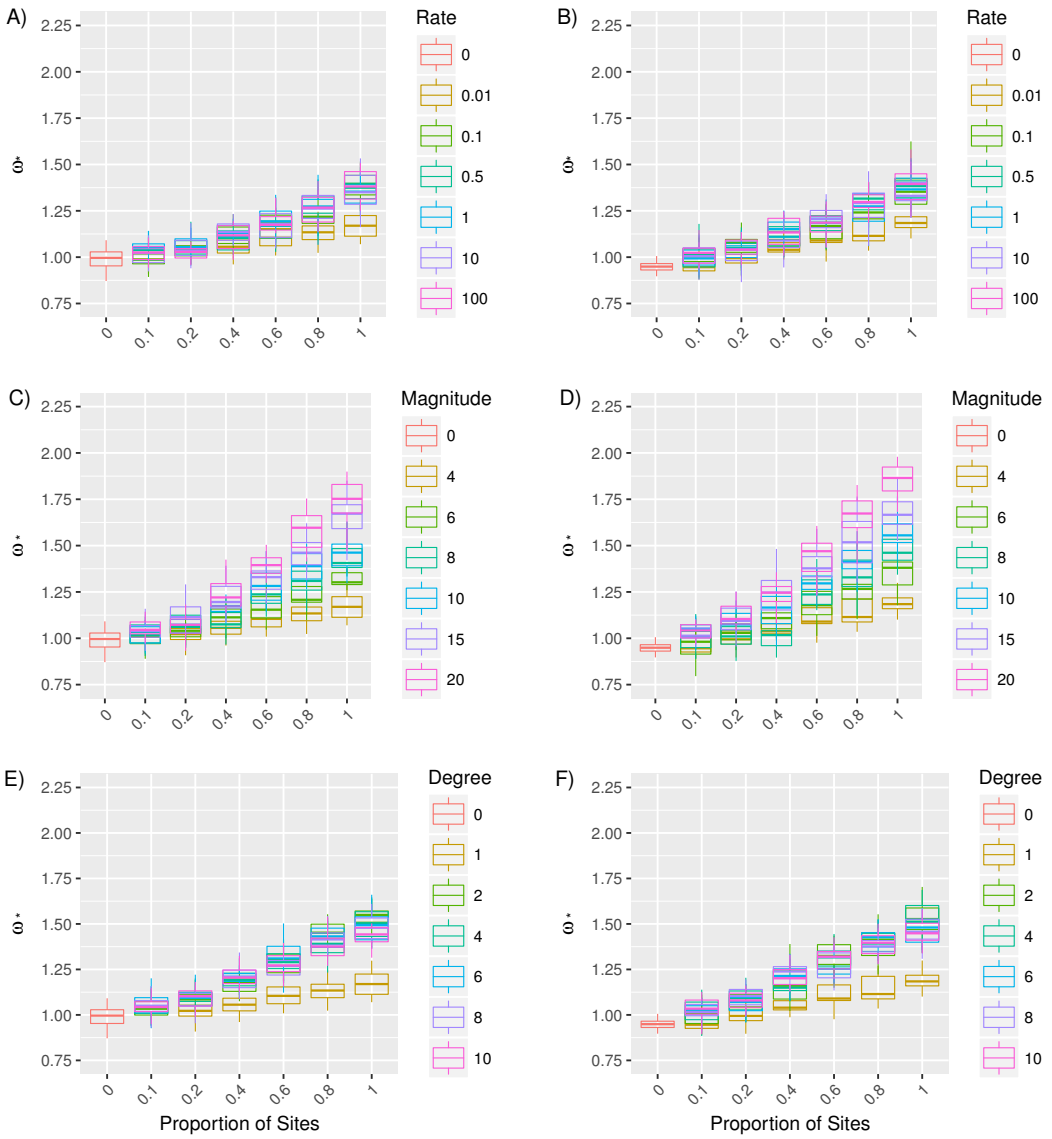


Figure 1 Simulations of an adaptive fitness landscape were performed using amino acid derived from the BRCA1 gene (Rodrigue and Lartillot, 2017) (A, C, E) and the nuclear concatenation by Lartillot and Delsuc (2012) (B, D, F). Varying the proportion of sites in combination with the magnitude of change had the greatest impact on ω_* under both sets of amino acid profiles (C,D). Increasing the rate of the Red Queen and the degree (the number of amino acid fitness values changed) eventually leads to a plateau effect (A,B, E,F).

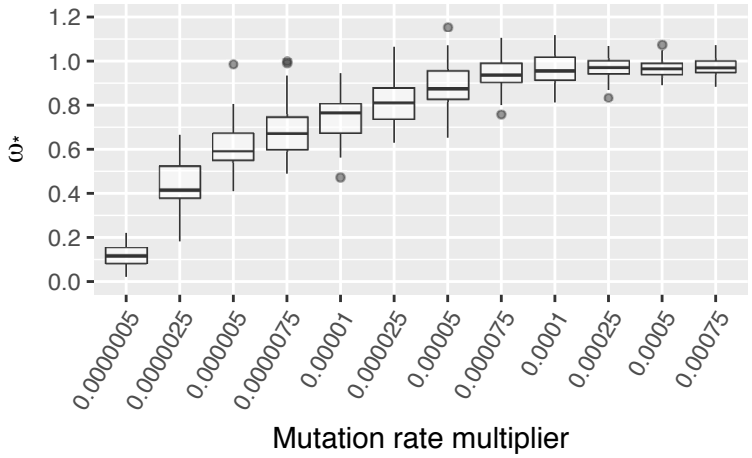


Figure 2 The effect of overall mutation rate on inferred values of ω_*.

Reproducing the results of Rodrigue and Lartillot (2017), we varied the *rate* of the Red-Queen (Figure 1A,B). When simulating alignments with 10% to 20% of sites under increasing rates of evolution, simulations were detected to be under adaptive evolution ($p(\omega_* > 1 \mid D) \geq 0.95$) with a frequency less than or equal to 30% of replicates. At least 40% of sites were required to be under the Red-Queen regime in conjunction with a moderate rate of change, to find adaptive evolution in more than half of the simulations. When increasing the rate of change for the evolutionary regime there appears to be a plateau effect, where subsequent increases to the rate have no further impact on ω_* (Figure 1A,B). In other words, increasing the rate of the Red Queen eventually leads to a saturation effect: if the fitness profiles are altered at a rate greater than the substitution rate, there is not sufficient time between Red Queen changes for the sequence to evolve toward the intermediate fitness optima, and the Red Queen starts "spinning its wheels".

The highest values of ω_* were reached by increasing the magnitude of changes in amino acid profiles, when the proportion of Red-Queen sites was high (Figure 1C,D). Simply stated, more drastic changes to amino acid profiles will increase the nonsynonymous flux in sequence evolution, leading to higher values of ω_*.

Altering the number of pairs of amino acid fitness values subject to change at a site led to a similar pattern observed with rate changes (Figure 1E,F). Examining simulations where only 10% or 20% of the sites in the alignment are considered, the proportion of simulations resulting in inferences with $\omega_* > 1$ shows little variation. However, beyond 40% of Red-Queen sites in the alignment, there is a rapid jump to greater than 90% of simulations being detected as being under adaptive evolution ($p(\omega_* > 1 \mid D) \geq 0.95$). However, with more than 40% of Red-Queen sites, after the increase in degree from one pair of amino acids to two, subsequent increases beyond two pairs have minimal apparent effect on on the value of ω_*. This can be understood from the fact that many changes to amino acid profiles will have little impact on the evolving sequence, if those changes are made to amino acid states that are not accessible via point mutation, which will tend to happen increasingly when changing multiple pairs of values.

The simulations described above were done with the objective of recreating adaptive evolution, through changes over time to the parameters controlling the amino acid fitness landscape. The next set of simulations was aimed at studying the model's behaviour with

pure mutation-selection simulations, where the overall amount of evolutionary signal is controlled by modulating the underlying mutation rate over time. We varied a multiplicative parameter to the mutation matrix given by Rodrigue and Lartillot (2017) from 0.000025 to 0.00075, again with 20 replicates for each setting. The results of the posterior mean ω_* across replicates are displayed with box-plots at each set of simulation conditions in Figure 2.

We see from Figure 2 that when the mutation rate is sufficiently high, the simulations tend to lead to ω_* around 1. When the mutation rate is low, however, ω_* is below one, and can be very low when the mutation rate is sufficiently low. Under these simulations, many sites have only a few substitutions (or none at all), such that when an analysis is conducted on the resulting artificial data sets, the Dirichlet process has very little evolutionary signal from which to infer the distribution of amino acid profiles across sites; loosely speaking, the model "chooses" to dispense with capturing purifying selection with the Dirichlet process apparatus (by adopting a low-dimensional, nearly flat configuration). In other words, when the overall evolutionary signal is very weak, the model given in Equation 18 reverts back to the simpler MG model given in (2), with ω_* effectively playing the role of ω.

6 Comparison of mutation-selection and classical frameworks on real data

The MutSelDP-ω_* model has only been applied to a handful of real data sets. To gain a broader empirical view of how the model reacts to real data sets, we analyzed a random sub-set of 4464 alignments taken from the OrthoMaM database (v8, Douzery et al., 2014). To get a general sense of the values of ω obtained across these alignments, we sorted the posterior mean values obtained and plotted them in Figure 3. Of the genes examined, 8.7% (388 genes) had 95% credibility intervals with ω_* greater than 1 (note that this implies that $p(\omega_* > 1 \mid D) > 0.975$). This is not entirely unexpected, as the majority of genes are not likely to be subjected to an ongoing Red-Queen over the mammalian phylogeny. What is somewhat surprising, however, is the extent to which most genes lead to ω_* well below 1. While is has been shown that epistasis can lead to ω_* values below 1 (Rodrigue and Lartillot, 2017), we suspect that many of these data sets are analogous to those simulated with very low mutation rates: they are highly conserved, and generally imply very few nonsynonymous substitutions. In other words, most alignments do not have sufficient evolutionary signal to reliably infer the Dirichlet process on amino acid profiles, such that the model favors a more compact means of fitting low nonsynonymous rates with low ω_* values, rather than several highly peaked amino acid profiles.

The same datasets from OrthoMaM were examined with CODEML (Yang, 2007), under the M7 and M8 models. Based on CODEML, 1301 genes reject the null M7 model in an LRT against M8 ($p < 0.05$). Among these, 497 genes had no sites within the gene detected to be under adaptive evolution as defined by $p(\omega > 1 \mid D) \geq 0.95$ with the Bayes Empirical Bayes approach. Here, we make the distinction between two classes of genes detected to be under adaptive evolution with M8: those that reject M7 in the LRT against M8, and those that also have at least one significant site ($p(\omega > 1 \mid D) \geq 0.95$). We compared genes in the latter class with the genes uncovered with the MutSelDP-ω_* model. There is a 70.1% overlap in the genes identified using the two methods, and this overlap climes to 82.5% if considering only genes with a length greater than 500 codons. Thus, it appears that the two methods are at least partially capturing the same features, but doing so in very different ways. It is particularly noteworthy that the MutSelDP-ω_* model is detecting adaptive regimes globally over the gene, whereas the M8 model has the potential for site-heterogeneous detection.

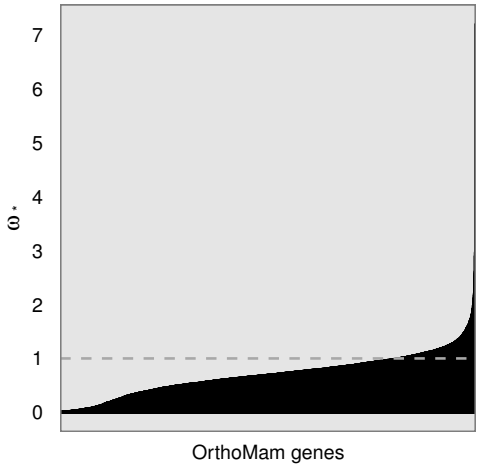


Figure 3 Analysis of a random subset of 4464 genes from the OrthoMam v8 database with the MutSelDP-ω_* model.

7 Conclusion

Although MutSelDP-ω_* is one of the richest codon substitution models proposed to date, its parameterization for detecting adaptive evolution is a simple univariate multiplier (ω_*) on nonsynonymous rates. Yet, it is already capable of detecting subtle instances of adaptive regimes in simulations, and has the potential to detect similar signals of adaptation in real data as the classical models with distributions of nonsynonymous rate multipliers.

It would of course be of great interest to expand the MutSelDP-ω_* model to allow for a distribution of ω_* values across sites, and/or across branches, in the same spirit as has been explored over the last few decades with classical codon models. Such models with multiple independent types of heterogeneity (e.g., distributions of amino acid profiles and distributions of ω_*) pose significant challenges, not the least of which will be their computational burden. Some short-cuts, such as utilizing empirically derived mixtures of amino acid profiles, along with a preset grid of ω_* values, could be worth considering for more speedy first-pass analyses of large data sets.

It would also be of interest to explore more simulations, incorporating more known features of the evolutionary process into the data-generating model. One glaring model violation in real data, laid bare in equation (10), is the assumption of a time-homogeneous effective population size (N_e). Understanding how the MutSelDP-ω_* model, or its eventual extensions, reacts to data simulated with changing effective population size over the tree would be an important first step. Other recent simulations (Laurin-Lemay et al., 2018a) have shown the CpG hypermutability can mislead some codon substitution models into detecting selection on synonymous codon usage. More generally, applying simulations to assessing the effects of these model violations, as well as others (Venkat et al., 2018; Jones et al., 2018), on mutation-selection-based models are in order to more carefully calibrate the reliability of the inferences to which they lead.

Finally, long-term modeling objectives should probably seek to integrate recent innovations into a single model, that could accommodate features such as uneven codon usage (e.g., as in Pouyet et al., 2016), variable effective populations size across the phylogeny, context-

dependent mutation rates, and epistatic effects (both within and across genes). Preliminary works exist that lay out the technical means of pursuing such a project, including ideas from Approximate Bayesian Computation (Laurin-Lemay et al., 2018b), and nested-MCMC systems (Robinson et al., 2003; Rodrigue et al., 2009; Kleinman et al., 2010; Rodrigue et al., 2010a). Bringing these ideas together could enable a framework for building models that are progressively more faithful to modern biological understanding of molecular evolution.

References

Anisimova, M. (2012). Parametric models of codon substitution. In Cannarozzi, G. M. and Schneider, A., editors, *Codon Evolution*, pages 12–33. Oxford University Press.

Bloom, J. (2017). Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biol. Direct*, 12:1.

Douzery, E. J., Scornavacca, C., Romiguier, J., Belkhir, K., Galtier, N., Delsuc, F., and Ranwez, V. (2014). Orthomam v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. *Mol. Biol. Evol.*, 31(7):1923–1928.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376.

Gojobori, T. (1983). Codon substitution in evolution and the saturation of synonymous changes. *Genetics*, 105(4):1011–1027.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11(5):725–736.

Guindon, S., Rodrigo, A. G., Dyer, K. A., and Huelsenbeck, J. P. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proc. Natl. Acad. Sci. USA*, 101(35):12957–12962.

Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15:910–917.

Holder, M. T., Zwickl, D. J., and Dessimoz, C. (2008). Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Phil. Tran. R. Soc. B*, 363:4013–4021.

Huelsenbeck, J. P. and Dyer, K. A. (2004). Bayesian estimation of positively selected sites. *J. Mol. Evol.*, 58:661–672.

Huelsenbeck, J. P., Jain, S., Frost, S. W. D., and Pond, S. L. K. (2006). A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl. Acad. Sci. USA*, 103:6263–6268.

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. (2018). Phenomenological load on model parameters can lead to false biological conclusions. *Molecular biology and evolution*, 35(6):1473–1488.

Kazmi, S. O. and Rodrigue, N. (2019). Detecting amino acid preference shifts with codon-level mutation-selection mixture models. *BMC Evol. Biol.*, 19:62.

Kleinman, C. L., Rodrigue, N., Lartillot, N., and Philippe, H. (2010). Statistical potentials for improved structurally constrained evolutionary models. *Mol. Biol. Evol.*, 27(7):1546–1560.

Lanave, C., Preparata, G., Saccone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, 20:86–93.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *J. Comput. Biol.*, 13:1701–1722.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. and Delsuc, F. (2012). Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, 66(6):1773–1787.

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes-MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, 62:611–615.

Laurin-Lemay, S., Philippe, H., and Rodrigue, N. (2018a). Multiple factors confounding phylogenetic detection of selection on codon usage. *Mol. Biol. Evol.*, 35(6):1463–1472.

Laurin-Lemay, S., Rodrigue, N., Lartillot, N., and Philippe, H. (2018b). Conditional approximate bayesian computation, a new approach for across-site dependency in high-dimensional mutation-selection models. *Mol. Biol. Evol.*, in press.

Li, W.-H., Wu, C.-I., and Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, 2(2):150–174.

Miyata, T. and Yasunaga, T. (1980). Molecular evolution of mrna: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *J. Mol. Evol.*, 16:23–36.

Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11(5):715–724.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, 3(5):418–426.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–936.

Nielsen, R. and Yang, Z. (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mtDNA. *Mol. Biol. Evol.*, 20:1231–1239.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. (1980). The evolution of genes: the chicken preproinsulin gene. *Cell*, 20(2):555–566.

Pond, S. K., Delport, W., Muse, S. V., and Scheffler, K. (2010). Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One*, 5(7):e11230.

Pouyet, F., Bailly-Bechet, M., Mouchiroud, D., and Guéguen, L. (2016). SENCA: A Multilayered Codon Model to Study the Origins and Dynamics of Codon Usage. *Gen. Biol. Evol.*, 8:2427–2441.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., and Thorne, J. L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.*, 18:1692–1704.

Rodrigue, N. (2013). On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics*, 193:557–564.

Rodrigue, N. and Aris-Brosou, S. (2011). Fast Baysian choice of phylogenetic models: prospecting data augmentation-based thermodynamic integration. *Syst. Biol.*, 60:881–887.

Rodrigue, N., Kleinman, C., Philippe, H., and Lartillot, N. (2009). Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codon. *Mol. Biol. Evol.*, 26:1663–1676.

Rodrigue, N. and Lartillot, N. (2014). Site-heterogeneous mutation-selection models within the phylobayes-mpi package. *Bioinformatics*, 30(7):1020–1021.

Rodrigue, N. and Lartillot, N. (2017). Detecting adaptation in protein-coding genes using a Bayesian site-heterogeneous mutation-selection codon substitution model. *Mol. Biol. Evol.*, 34(1):204–214.

Rodrigue, N., Philippe, H., and Lartillot, N. (2008a). Bayesian comparisons of codon substitution models. *Genetics*, 180:1579–1591.

Rodrigue, N., Philippe, H., and Lartillot, N. (2008b). Uniformization for sampling realizations of markov processes: applications to bayesian implementations of codon substitution models. *Bioinformatics*, 24(1):56–62.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010a). Mechanistic revisions of phenomenological modeling strategies in molecular evolution. *Trends Genet.*, 26:248–252.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010b). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. USA*, 107(10):4629–4634.

Seo, T. K. and Kishino, H. (2008). Synonymous Substitutions Substantially Improve Evolutionary Inference from Highly Diverged Proteins. *Syst. Biol.*, 57:367–377.

Spielman, S. J. and Wilke, C. O. (2015). The relationship between dN/dS and scaled selection coefficients. *Mol. Biol. Evol.*, 32(4):1097–1108.

Tamuri, A. U., dos Reis, M., and Goldstein, R. A. (2012). Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190:1101–115.

Tamuri, A. U., Goldman, N., and dos Reis, M. (2014). A Penalized Likelihood Method for Estimating the Distributionof Selection Coefficients from Phylogenetic Data. *Genetics*, 197:257–271.

Venkat, A., Hahn, M. W., and Thornton, J. W. (2018). Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nature ecology & evolution*, 2(8):1280.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39:306–14.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxfors Series in Ecology and Evolution.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24(8):1586–1591.

Yang, Z. and Dos Reis, M. (2010). Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution*, 28(3):1217–1228.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, 19(6):908–917.

Yang, Z. and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.*, 25:568–579.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449.

Yang, Z., Wong, W. S., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, 22(4):1107–1118.

Chapter 4.6 The Nature and Phylogenomic Impact of Sequence Convergence

Zhengting Zou
Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA
ztzou@umich.edu
https://orcid.org/0000-0003-1716-5090

Jianzhi Zhang
Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA
jianzhi@umich.edu
https://orcid.org/0000-0001-6141-1290

Abstract

Protein sequence convergence refers to substitutions leading to the same amino acid residue at the same position of a protein in multiple independent evolutionary lineages. Protein sequence convergence is often viewed as adaptive signal so is of great interest to evolutionary biologists. In this article, we review complications in identifying sequence convergences, statistical tests of the null hypothesis that the observed convergence events in a protein are attributable to chance alone, interpretations of genome-wide observations of sequence convergence, and a comparison in the susceptibility of molecular and morphological characters to convergence and its phylogenetic implications. We highlight the substantial progresses made in the last two decades and point out the main challenges at the present.

How to cite: Zhengting Zou and Jianzhi Zhang (2020). The Nature and Phylogenomic Impact of Sequence Convergence. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 4.6, pp. 4.6:1–4.6:17. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 Introduction

Convergent evolution, or simply convergence, refers to the independent emergences of the same state of a character in two or more lineages of living organisms. Well-known examples of convergence include the origins of camera-type eyes in cephalopods and vertebrates, and the emergences of wings from forelimbs in birds and bats. Evolutionary biologists are interested in convergence primarily for three reasons. First, because complex characteristics such as camera-type eyes and wings are unlikely to have emerged more than once simply by chance, convergent evolution of complex characteristics is believed to reflect similar adaptations in multiple lineages. Second, convergence indicates that evolution is predictable to some extent, either because there are few viable solutions to a problem or the best solutions are similar in different lineages. Third, convergence confuses phylogenetic analysis, because true phylogenetic signals are based on identity by descent, which, however, is not easy to distinguish from false signals of identity by convergence.

The study of convergence has a long history. Convergence was already discussed in Darwin's Origin of Species as "analogical resemblances"; examples mentioned included body shape and fin-like forelimb of dugongs and whales, morphological resemblance between

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 4.6; pp. 4.6:1–4.6:17

A book completely handled by researchers.
No publisher has been paid.

European and Asian domestic pigs, and electric organs in different fish lineages (Darwin, 1859). Numerous morphological and functional convergences have since been reported, such as similar web architectures of spiders occupying the same habitat type on different Hawaiian islands (Blackledge and Gillespie, 2004), similar bill shape shifts of tidal marsh sparrows in North America (Grenier and Greenberg, 2005), morphological similarities among trunk-ground dwelling anoles on multiple Greater Antillean islands (Langerhans et al., 2006), and intercontinental pairs of desert iguana species with matching habitats (Melville et al., 2006). The list goes on and on with examples across virtually the whole tree of life (Nevo, 1979; Moore and Willmer, 1997; Wittkopp et al., 2003; Fong et al., 2005; Maruyama and Parker, 2017). As our understanding of biology progresses to the molecular level, convergence has also been discovered at the level of molecular phenotypes. For example, the independently evolved pATOM36 protein and MIM complex serve as importers on the mitochondrial outer membrane in trypanosomes and yeast, respectively (Vitali et al., 2018), and rhodopsins of vertebrates and of the visually competent box jellyfish have acquired similar tertiary structures to enable high-fidelity photoreception (Gerrard et al., 2018).

Convergence can occur not only at the phenotypic level but also at the molecular genetic level (Stewart et al., 1987; Doolittle, 1994; Arendt and Reznick, 2008; Manceau et al., 2010). This can include, for example, amino acid sequence changes at different sites of the same protein across multiple lineages (Protas et al., 2006; Rosenblum et al., 2010; Linnen et al., 2013; Zhou et al., 2015; Chikina et al., 2016), or independent formations of a chromosomal cluster of the same set of genes via relocation (Slot and Rokas, 2010). The most studied convergence at the molecular genetic level is, however, sequence convergence. Formally, sequence convergence is defined by independent changes leading to the same nucleotide or amino acid residue at the corresponding sequence positions in multiple lineages. Sequence convergence is often divided into parallel changes and convergent changes, depending on whether the ancestral states prior to the changes are the same or differ among the lineages (Zhang and Kumar, 1997). Hereinafter, we collectively refer to these two types as convergence unless otherwise mentioned. With the rapid accumulation of genome sequences from a variety of organisms, recent years have seen a surge in the report of sequence convergence, prompting a series of questions about the prevalence, adaptiveness, and phylogenetic impacts of sequence convergence. There have also been developments of methods to test whether sequence convergence is attributable to chance. We discuss these aspects of progress in this review.

2 Tests of adaptive sequence convergence in individual genes

Because phenotypic convergences are commonly viewed as strong indications of adaptive evolution, sequence convergences tend to be viewed similarly. However, because there are only four possible states at a nucleotide position and 20 possible states at an amino acid position, and because many of these states are not selectively allowed, the actual number of states permitted per nucleotide or amino acid position is quite small. This makes it possible for sequence convergence to occur simply by chance via neutral evolution instead of by a common selective force. Zhang and Kumar (1997) pioneered the modeling of chance sequence convergence. They proposed a test to examine whether the observed number of parallel or convergent amino acid substitutions is attributable to chance alone. Zou and Zhang (2015a) improved the test by considering different amino acid equilibrium frequencies at different sites, making the neutral model more realistic and the test more reliable. Below we briefly describe Zou and Zhang's test.

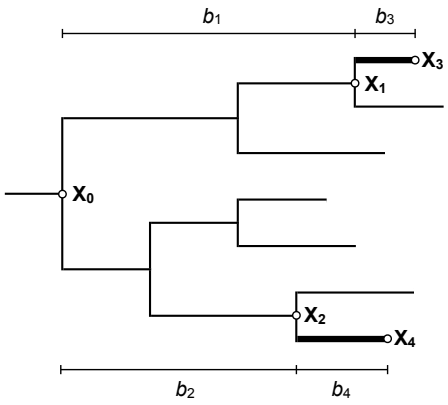


Figure 1 Counting the observed and expected numbers of events of sequence convergence between two branches on a tree. At a given position, the amino acids at nodes X_0 – X_4 are x_0 – x_4, respectively. Thick branches correspond to the converging lineages. The relevant branch lengths are indicated by the b values.

Let us take amino acid sequence evolution as an example and consider the possibility of sequence convergence between two focal branches (X_1 to X_3 and X_2 to X_4, respectively) of an arbitrary phylogeny shown in Figure 1. Let x_1, x_2, x_3, and x_4 be the amino acids at nodes X_1, X_2, X_3, and X_4, respectively. Convergent changes at a site can be defined by $x_1 \neq x_3$, $x_2 \neq x_4$, $x_1 \neq x_2$, and $x_3 = x_4$, whereas parallel changes can be defined by $x_1 \neq x_3$, $x_2 \neq x_4$, $x_1 = x_2$, and $x_3 = x_4$. Starting from an arbitrary root node X_0 with its state x_0, the probability of observing any conformation $X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4$ can be calculated based on a designated Markovian model of protein sequence evolution by $P(x_1, x_2, x_3, x_4 \mid x_0) = P(x_1 \mid x_0)P(x_3 \mid x_1)P(x_2 \mid x_0)P(x_4 \mid x_2)$. Based on the tradition of modelling sequence evolution as a continuous-time Markov process (see Chapter 1.1 [Pupko and Mayrose 2020]), we can calculate each conditional distribution by $P(Y \mid X = x) = I^{(x)}M^b$. Here $I^{(x)}$ is an indicator vector of size 20, with the element corresponding to amino acid x being 1 and all other elements being 0, M is the substitution matrix such as the empirical JTT matrix (Jones et al., 1992) or WAG matrix (Whelan and Goldman, 2001), and b is the branch length between node X and Y. Thus, $P(Y = y \mid X = x)$ is the probability of observing amino acid state y at node Y given state x at node X. The probabilities of parallel and convergent substitutions can thus be calculated by:

$$P_{\text{parallel}} = \sum_{x_1 \neq x_3, x_2 \neq x_4, x_1 = x_2, x_3 = x_4} P(x_1, x_2, x_3, x_4 \mid X_0 = x_0)P(X_0 = x_0)$$

$$P_{\text{convergent}} = \sum_{x_1 \neq x_3, x_2 \neq x_4, x_1 \neq x_2, x_3 = x_4} P(x_1, x_2, x_3, x_4 \mid X_0 = x_0)P(X_0 = x_0)$$

Here $P(X_0 = x_0)$ can be obtained from an indicator vector according to the inferred ancestral state (Zou and Zhang, 2015a). We can further compute $P_{\text{convergence}} = P_{\text{convergent}} + P_{\text{parallel}}$. For a pair of focal branches, such a probability of convergence can be calculated for each amino acid site. Because $P_{\text{convergence}}$ is small, the number of convergences at each site can be modeled by a Poisson distribution with $P_{\text{convergence}}$ as the expectation. So, the number of convergence events for an amino acid sequence is a new Poisson random variable with the expectation equal to the sum of $P_{\text{convergence}}$ across all sites. Given this expectation, one can compute the probability of occurrence of the observed number or more convergence events

from the upper tail probability of the corresponding Poisson distribution (Zou and Zhang, 2015a). A significant test result indicates that the observed convergence events are not fully attributable to chance, implying the involvement of a common selective force in the multiple lineages considered. Similarly, one can also separately test whether the observed parallel changes and convergent changes are more numerous than expected by chance, using P_{parallel} and $P_{\text{convergent}}$, respectively (Zou and Zhang, 2015a).

Applying the above statistical test of convergence or its original version (Zhang and Kumar, 1997) has revealed a number of cases of sequence convergence that are unattributable to chance alone. For example, the prestin protein is a member of the SLC26 anion-transport family providing the electromobility of outer hair cells thought to be responsible for cochlear amplification, an active process that confers sensitivity and frequency selectivity to the mammalian auditory system. Prestin showed significantly more parallel substitutions than expected in the origins of echolocating bats and toothed whales, two mammalian groups that independently evolved echolocation (Li et al., 2010; Liu et al., 2010). Given prestin function[1], this observation suggests prestin contribution, especially by the parallel substitutions, to the evolution of echolocation. Indeed, subsequent cell functional assays showed that the replacement of amino acid N with T at position 7, one of the parallel substitutions observed, converts the prestin of a non-echolocator to that of an echolocator in a biophysical property associated with mammalian high-frequency hearing (Liu et al., 2014). There were also reports of sequence convergence beyond what chance can explain between two lineages of echolocating bats that are thought to have independently evolved echolocation, although the functional role of the sequence convergence has yet to be demonstrated (Liu et al., 2011). More parallel amino acid substitutions than expected by chance were also observed in the independent evolution of a digestive ribonuclease in African and Asian leaf-eating monkeys, and *in vitro* assays showed that the parallel substitutions are responsible for the parallel improvements of the enzyme's activity in the two lineages (Zhang et al., 2002; Zhang, 2006).

Some studies experimentally demonstrated the functional role of sequence convergence without formally testing whether the convergent/parallel substitutions are attributable to chance. This was shown, for instance, for the role of sequence convergence in the elevated oxygen affinity of the hemoglobin in independently adapted hummingbirds in the Andes (Projecto-Garcia et al., 2013). In addition, there are many cases of sequence convergence in lineages that have experienced phenotypic convergence, but neither the causal relation between sequence convergence and phenotypic convergence nor the implausibility of sequence convergence by chance has been established (Christin et al., 2008; Jost et al., 2008; Shen et al., 2010; Feldman et al., 2012; Zhen et al., 2012; Ujvari et al., 2015).

Four criteria have been proposed to establish adaptive parallel/convergent evolution at the protein sequence level (Zhang, 2006). First, similar changes in protein function occur in independent evolutionary lineages. Second, parallel/convergent amino acid substitutions are observed in these proteins. Third, the parallel/convergent substitutions are not attributable to chance alone and therefore must have been driven by a common selective pressure. Fourth, the parallel/convergent substitutions are responsible for the parallel functional changes. These criteria are more stringent than simply showing a significant result from the statistical test mentioned (criterion 3), because one should know the protein functional consequence of the sequence convergence and the selective agent when claiming adaptation. Most claims of adaptive sequence convergence are based on criterion 2 only, some also satisfy criteria 1 and/or 3, but very few satisfy all four criteria.

[1] See Chapter 4.2 (Robinson-Rechavi 2020) for a review of how gene functions are defined.

To detect sequence convergence, one must first know the phylogenetic relationships of the sequences concerned. In theory, the tree used should be the gene tree instead of the species tree, but most people use the species tree instead, probably because the estimation of the gene tree is less reliable than that of species tree, which can be inferred using many genes together. When the species tree is used, some inferred sequence convergences may be false positives due to the discordance between the gene tree and species tree (Mendes et al., 2016). Thus, it is important to ensure that the underlying tree of the sequence evolution is correctly assumed in the study of sequence convergence.

3 Genomic patterns of sequence convergence: adaptive or neutral?

The abundance of genome sequence data now allows researchers to discover sequence convergence at the genomic scale. Without finding the specific reason of convergence at individual sites, one can ask whether the total number of convergence events observed in a genome exceeds the neutral expectation. Four approaches have been used to address this question. The first is the statistic $\Delta SSLS$ (difference in site-specific likelihood support), which is the difference in log likelihood value of a site under two alternative tree topologies. For example, Liu et al. (2011) evaluated each nucleotide site of the mitochondrial genome sequence alignment containing squamate reptile species for its likelihood support of a widely accepted nuclear tree topology versus a radically different mitochondrial topology splitting the supposedly monophyletic Iguania. It was found that most sites support the nuclear tree, while a small number of sites strongly favor the convergent mitochondrial topology. The log likelihood nature of $\Delta SSLS$ allows summation over sites to derive a gene-specific statistic (Parker et al., 2013). However, there is no explicit neutral expectation of $\Delta SSLS$ for a site or gene; consequently, the $\Delta SSLS$ distribution for different sites or genes is empirical. One can identify sites or genes in each tail of the $\Delta SSLS$ distribution, but cannot prove based on this information whether the convergence signals of these sites or genes are due to positive selection, because any distribution has tails (Zou and Zhang, 2015b). Therefore, while this method allows identifying sites/genes with the strongest convergence signals, it does not allow testing whether the observed convergence signals result from positive selection.

The second approach is to use observed rates of sequence divergence as a control when studying sequence convergence. A divergence event at a site between two branches refers to independent substitutions at the site in the two branches resulting in different nucleotide or amino acid states. The numbers of convergence (Cv) and divergence (Dv) events for a pair of branches are strongly correlated such that the $\frac{Cv}{Dv}$ ratio typically does not vary greatly among different pairs of branches (Castoe et al., 2009; Thomas and Hahn, 2015). A significantly higher $\frac{Cv}{Dv}$ ratio for a focal branch pair relative to other branch pairs would suggest a deviation in the focal branches. However, because the $\frac{Cv}{Dv}$ ratio under neutral evolution is unknown, one cannot prove that the deviation results from adaptive convergence in the focal branches. Furthermore, recent studies showed that, even under neutral evolution, the $\frac{Cv}{Dv}$ ratio decreases with the divergence between the branches concerned (Goldstein et al., 2015; Zou and Zhang, 2017), violating the assumption that the $\frac{Cv}{Dv}$ ratio is expected to be constant among all pairs of branches. However, one can classify both convergence and divergence events into two types: substitutions starting from the same state and those starting from different states. The two types of convergence events are precisely parallel and convergent substitutions, respectively. The convergent $\frac{Cv}{Dv}$ ratio and parallel $\frac{Cv}{Dv}$ ratio are each expected to be constant irrespective of the divergence between the branches (Zou and Zhang, 2017).

The third approach to testing adaptive convergence between a pair of branches is to use comparable control branch pairs (Projecto-Garcia et al., 2013; Foote et al., 2015; Zou and Zhang, 2015b; Natarajan et al., 2016; Xu et al., 2017). For instance, to test whether there is an excess in sequence convergence between echolocating bats and dolphins (focal branch pair), one could compare the focal branch pair with the control branch pair of echolocating bats and the cow, which represents a (non-echolocating) sister lineage of dolphins (Zou and Zhang, 2015b). Interestingly, in this case, the focal branch pair has fewer sequence convergences than the control branch pair (Zou and Zhang, 2015b). This comparison can be further controlled by the number of divergence events in the two branch pairs, accounting for potential differences in branch lengths. That is, one can construct a contingency table with Cv and Dv values of the two branch pairs, which can be statistically compared by a G-test. Notably, because sister taxa can have different branch lengths, the expected $\frac{Cv}{Dv}$ ratio is only equal when the convergent $\frac{Cv}{Dv}$ and divergent $\frac{Cv}{Dv}$ are separately considered, as mentioned above. Recently, Xu et al. (2017) applied a more stringent criterion in counting sequence convergence events in order to increase the probability of identifying adaptive sequence convergences. They compared three mangrove species with their respective non-mangrove sister species as well as a species that is the outgroup of all six species, and counted a convergence event at a site only when all mangrove species share the same amino acid state that differs from the amino acid state conserved among all four non-mangrove species. Their simulation showed that applying this criterion of convergence at conserved sites (CCS) substantially reduces chance convergence or convergence due to incorrect inference of ancestral states (Xu et al., 2017). Nevertheless, not all CCS events are necessarily adaptive, and Xu et al. (2017) selected candidate genes for adaptive convergence according to the number of CCS events per gene, based on an arbitrary cutoff. Thus, the CCS method provides candidates for adaptive convergence rather than proving adaptive convergence.

None of the above three approaches estimate the number of convergence events expected under neutral evolution. Consequently, they can compare the amount of sequence convergence among genes or among branch pairs, but cannot tell whether the amount of convergence observed exceeds the neutral expectation. The fourth and final approach differs from the above approaches in that it compares the observed amount of convergence with the neutral expectation. The neutral expectation is estimated by conducting computer simulations of sequence evolution or is probabilistically calculated. For instance, Rokas and Carroll (2008) used simulations to generate sequence alignments of the same size as the real data, using relatively simple models whose parameters are estimated from the actual data. They then regarded the number of convergence events observed from the simulated alignments as the neutral expectation. Zou and Zhang (2015a) directly calculated the expected number of convergence events between focal branch pairs as described in the previous section. Regardless of the method used in deriving the neutral expectation, the key is the substitution model and its parameters, because using different models or parameters results in drastically different neutral expectations (Zou and Zhang, 2015a). Rokas and Carroll (2008) reported that the number of convergence events observed at the genomic scale is much greater than the neutral expectation and suggested that this excess may be due to positive selection. However, in estimating the neutral expectation, they assumed equal amino acid compositions across sites, which is unrealistic and may lead to underestimation of the neutral expectation. In a subsequent study, Zou and Zhang (2015a) showed that the amount of sequence convergence observed at the genomic scale is compatible with neutral expectations derived under realistic substitution models. Below we summarize their analyses and results.

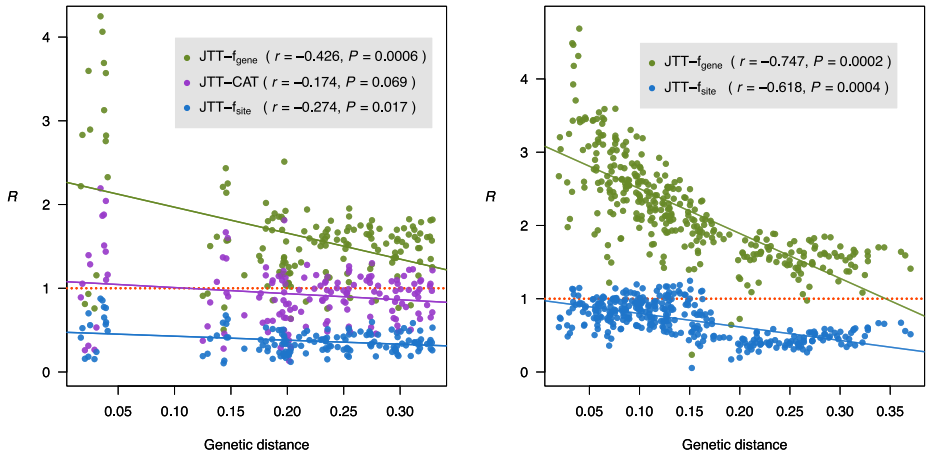
In 5,935 orthologous protein alignments of 12 *Drosophila* species, totaling 2,028,428

amino acid sites after the removal of gaps and ambiguous sites, 650 and 292 sites respectively experienced parallel and convergent substitutions in the two exterior branches leading to *D. yakuba* and *D. mojavensis*. Are these observed numbers of sites with parallel and convergent substitutions significantly greater than the corresponding neutral expectations? Zou and Zhang (2015a) examined three different neutral models. The first is the gene-specific JTT-f_{gene} model, which is based on the average substitution patterns of many proteins (Jones et al., 1992) with the equilibrium frequencies of the 20 amino acids in the model replaced with the observed amino acid frequencies of the protein concerned. The second neutral model considered is the site-specific JTT-f_{site} model, in which the equilibrium amino acid frequencies are replaced with the observed amino acid frequencies at the site concerned across all sequences in the alignment. One caveat in applying the JTT-f_{site} model is that, because the number of taxa used is smaller than 20 and because the total branch length of the *Drosophila* tree is also much smaller than 20, the observation of a limited number of different amino acids at a site may not mean that only those observed amino acids are acceptable but could be due to insufficient evolutionary time and taxon sampling for all acceptable amino acids to appear. Zou and Zhang (2015a) thus tried a third neutral model, JTT-CAT (Lartillot and Philippe, 2004) to estimate the expected numbers of convergent and parallel sites. Instead of having one set of equilibrium amino acid frequencies for all sites of a protein (JTT-f_{gene}) or one set per site (JTT-f_{site}), CAT uses a Bayesian mixture model for among-site heterogeneities in amino acid frequencies (see Chapter 1.4 [Lartillot 2020]). It estimates the total number of classes of sites and their respective amino acid frequencies, as well as the affiliation of each site to a given class. Due to the computational intensity of parameter estimation under JTT-CAT, Zou and Zhang (2015a) analyzed 1,081 relatively long proteins from the entire set of 5,935 proteins in an attempt to acquire the most information with the least amount of computer time.

The expected numbers of convergent and parallel sites, as well as the ratios (R) of the observed to expected numbers, are presented in Table 1 under each of the three neutral models. One can see that R varies from significantly above 1 to significantly below 1 among different neutral models (Table 1).

Type of sites	Number of sites examined	Observed number of sites	Expected number of sites		R	P-value
			Substitution model	Number of sites		
Convergent sites	2,028,428	292	JTT-f_{gene}	194.2	1.50	3.8E-11
	2,028,428	292	JTT-f_{site}	475.2	0.61	9.4E-20
	780,615	93	JTT-CAT	118.0	0.79	1.0E-3
Parallel sites	2,028,428	650	JTT-f_{gene}	388.6	1.67	3.2E-34
	2,028,428	650	JTT-f_{site}	2125.7	0.31	8.8E-309
	780,615	218	JTT-CAT	184.8	1.18	9.4E-3

Table 1 Observed numbers of sites experiencing convergent and parallel substitutions and the corresponding numbers expected under various neutral models of amino acid substitution. Reprinted with permission from Zou and Zhang (2015a). Results presented are for the two exterior branches leading to *D. yakuba* and *D. mojavensis*, respectively. R is defined as the ratio between the observed number and expected number. For the computation of the P-value, a statistical test is conducted under the assumption that the number of convergent (or parallel) sites follows a Poisson distribution with the mean equal to the expected number. When the observed number is smaller than the expected, the lower tail probability is given; when the observed number is larger than the expected, the upper tail probability is given.



(a) (b)

Figure 2 Ratios (R) of observed numbers of molecular convergences to the expected numbers in the protein evolution of *Drosophila* and mammals. (a) Scatter plot showing R against the genetic distance between the two branches concerned in the phylogeny of 12 *Drosophila* species. The R values under JTT-f_{gene} and JTT-f_{site} are based on all 5,935 proteins, whereas those under JTT-CAT are based on a subset of 1,081 proteins. (b) Scatter plot showing R against the genetic distance between the two branches considered for 2,759 proteins in the phylogeny of 17 mammals. In both panels, each dot represents one branch pair, and different colors show the results under different neutral models. Genetic distance is the number of amino acid substitutions per site between the two younger ends of the two branches considered. Solid lines show linear regressions. The r values are Pearson's correlation coefficients. P values are from Mantel tests. The horizontal red dotted line shows $R = 1$. The figure was redrawn using data from Zou and Zhang (2015a).

Thus, the answer to the question of whether there are more sequence convergences than the chance expectation depends on the neutral model assumed. Similar patterns were found when other branch pairs in the *Drosophila* tree were examined (Figure 2(a)). Zou and Zhang (2015a) further repeated this analysis in a set of 17 mammals. The data comprised 2,759 one-to-one orthologous proteins, with a total length of 1,079,696 amino acid sites. While the large data size prohibited them from using JTT-CAT, the analysis showed that R tends to exceed 1 under JTT-f_{gene} but becomes close to or even smaller than 1 under JTT-f_{site} (Figure 2(b)). Because models considering among-site heterogeneity in equilibrium amino acid frequencies almost always fit actual protein sequences better than comparable models assuming among-site homogeneity (Lartillot and Philippe, 2004, 2006), the findings from using the fourth approach suggest that the observed sequence convergence at the genomic scale is generally explainable by chance.

Figure 2 also shows an interesting pattern that R decreases with the genetic distance (number of amino acid substitutions per site) between the two younger ends of the two branches considered. A similar trend was independently reported for vertebrate mitochondrial proteins (Goldstein et al., 2015). Two explanations have been proposed. First, incomplete lineage sorting could make a gene tree different from the species tree, causing false inferences of convergences under the species tree. When the genetic distance between the two lineages considered increases, such false positive errors are expected to reduce, resulting in a negative correlation between R and the genetic distance (Mendes et al., 2016). Second, due to interactions among amino acid residues, the amino acids acceptable at a

site may change in evolution as a result of substitutions at other sites, such that an amino acid allowed at a site in one part of a tree becomes prohibited in another part of the tree, reducing the probability of sequence convergence with the genetic distance. Because the neutral models considered here do not include this factor, the neutral expectation of convergence is presumably overestimated, and R underestimated, when the genetic distance is large (Zou and Zhang, 2015a). While empirical evidence for the first reason exists, further analysis after excluding this factor still shows a negative correlation between R and the genetic distance, suggesting that the second reason may also exist in the data analyzed (Zou and Zhang, 2017). Importantly, evolutionary shifts in amino acid compositions at a site were observed when large alignments of hundreds to thousands of orthologous proteins were examined (Zou and Zhang, 2015a), and case studies showed that the same amino acid substitution can sometimes cause different or even opposite functional effects in homologous proteins (Zhang, 2003; Natarajan et al., 2016).

Note that, in all of the above analyses, amino acid substitutions are assumed to follow the JTT matrix with a set of equilibrium amino acid frequencies that could vary among sites or proteins. Recent studies found that the JTT substitution matrix does not apply universally and that different species show species-specific, genome-wide substitution patterns (Zou and Zhang, 2019). This means that amino acid substitution patterns are more diverse across species than generally thought. Consequently, one should be cautious in interpreting results from using the JTT matrix when studying sequence convergence.

4 Convergence as noise in phylogenetics

When discussing "analogical resemblances", Darwin pointed out that such resemblances "will not reveal—will rather tend to conceal their blood-relationship", so are "almost valueless to the systematist" (Darwin, 1859). Convergence is actually worse than being valueless, because it confuses phylogenetic inference and should be removed in phylogenetics if at all possible. Traditionally, phylogenetic trees of different organisms are inferred using morphological, physiological, or behavioral characters, collectively referred to as morphological characters hereinafter. The advent of molecular biology, especially the accumulation of sequenced genomes, supplied numerous molecular characters in the form of DNA and protein sequences, which are often considered more suitable than morphological characters for phylogenetic inference (Jousselin et al., 2003; Perelman et al., 2011; Wake et al., 2011; Legg et al., 2013; Springer et al., 2013; Jarvis et al., 2014). A major reason for this consideration concerns convergence. Compared with morphological characters, molecular characters are believed by many to be less susceptible to convergence (Givnish and Sytsma, 1997; Page and Holmes, 1998; Jousselin et al., 2003; Gaubert et al., 2005; Wiens et al., 2010; Wake et al., 2011; Davalos et al., 2012; Legg et al., 2013; Springer et al., 2013; Davalos et al., 2014). Nevertheless, this belief appears to have arisen in the early days of molecular systematics when morphological convergence had long been known while molecular convergence had not. As mentioned above, recent genetic and genomic studies revealed a large number of convergence events in protein sequence evolution. Zou and Zhang (2016) therefore compared the two character types, focusing on a large dataset containing both morphological and molecular characters that was previously used for jointly inferring the mammalian species tree. The data consist of 3,414 parsimony informative morphological characters and 5,722 parsimony informative amino acid sites for 46 extant and 40 fossil species (O'Leary et al., 2013). Below we summarize the analyses and findings from Zou and Zhang (2016).

Identifying character convergence requires the correct phylogeny, but because the mam-

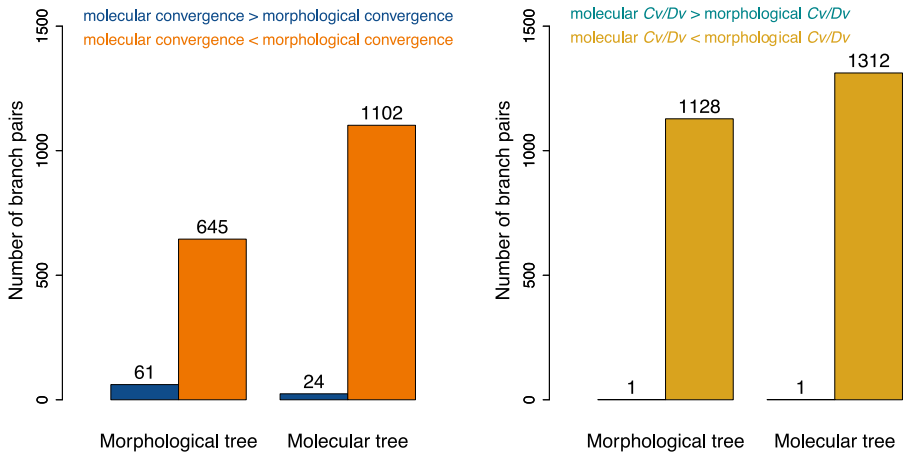
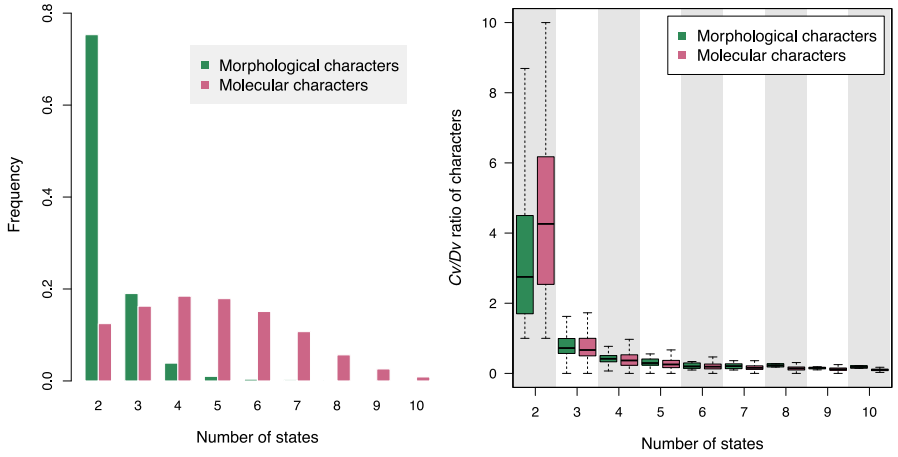


Figure 3 Comparison between morphological and molecular (sequence) convergences in mammalian evolution. (a) Comparison between the number of branch pairs for which the mean number of convergences per morphological character significantly exceeds that per molecular character (orange) and the number of branch pairs for which the number of convergences per molecular character significantly exceeds that per morphological character (blue) under each of two trees considered. (b) Comparison between the number of branch pairs for which the convergence/divergence ($\frac{Cv}{Dv}$) ratio is significantly greater for morphological characters than molecular characters (yellow) and the number of branch pairs for which $\frac{Cv}{Dv}$ is significantly lower for morphological characters than molecular characters (green) under each of two trees considered. In both panels, significance is defined by Q-value < 0.05. Number of branch pairs for a bar is indicated above the bar. The figure was redrawn using data from Zou and Zhang (2016).

malian tree is not completely resolved, Zou and Zhang (2016) considered three trees, respectively reconstructed using the morphological characters only, molecular characters only, and both types of characters in the data. Under each tree, they inferred the ancestral states at all interior nodes for each character by parsimony. For each pair of independent branches that can be investigated for convergence, they identified characters that showed convergence and compared the mean number of convergences per character between morphological and molecular characters. Among 3,396 investigated pairs of branches in the morphological tree, the number of branch pairs with a significantly higher number of convergences per morphological character than that per molecular charter substantially exceeds the number of branch pairs with a significantly lower number of convergences per morphological character than that per molecular character (Figure 3(a)). The mean number of convergence per morphological character is 1.7 times that per molecular character. When comparing the $\frac{Cv}{Dv}$ ratio introduced early, they also found morphological characters to exhibit overwhelmingly larger $\frac{Cv}{Dv}$, compared with molecular characters (Figure 3(b)). The mean $\frac{Cv}{Dv}$ ratio of morphological characters is 4.0 times that of molecular characters. When the above analyses were repeated under the molecular tree, even more convergences and higher $\frac{Cv}{Dv}$ ratios were found for morphological characters relative to those for molecular characters (Figure 3). Similar results were obtained under the total evidence tree.

Zou and Zhang (2016) noted that 75.2% of parsimony-informative morphological characters are binary in the data of O'Leary et al. (2013) (Figure 4(a)). Because binary characters can only have one kind of change given an ancestral state, it is obvious that they are susceptible to convergence once multiple changes occur. By contrast, only a small fraction (12.4%)



(a) (b)

Figure 4 Morphological characters tend to have fewer states than molecular characters. (a) Frequency distribution of the number of states per character. (b) $\frac{Cv}{Dv}$ ratio of a character decreases as the number of states increases. $\frac{Cv}{Dv}$ ratio of a character is the sum of convergences across all branch pairs divided by that of divergences. The top and bottom edges of a box represent the first and third quartiles of the distribution, respectively, while the thick line inside the box represents the median. The two whiskers show the maximum value not greater than the first quartile plus 1.5 times the box height and the minimum value not smaller than the third quartile minus 1.5 times the box height, respectively. $\frac{Cv}{Dv}$ ratios are calculated under the morphological tree. The same pattern is observed when $\frac{Cv}{Dv}$ ratios are calculated under the molecular tree. The figure was redrawn using data from Zou and Zhang (2016).

of molecular characters are binary (Figure 4(a)). The median number of states is five for molecular characters, significantly higher than that (two) for morphological characters ($P < 10^{-300}$). The probability of convergence relative to that of divergence for a character is expected to decrease with the number of states. Let the $\frac{Cv}{Dv}$ ratio of a character be the sum of Cv values across all branch pairs divided by the sum of Dv values across all branch pairs for the character. Indeed, the $\frac{Cv}{Dv}$ ratio decreases with the number of states for both types of characters (Figure 4(b)) and this trend remains after the control of evolutionary rate (represented by number of steps inferred on the tree). It was estimated that the $\frac{Cv}{Dv}$ ratio of an average morphological character is 0.89 times that of a molecular character with the same number of states. These results indicate that, compared with molecular characters, the higher convergence of morphological characters is caused by having fewer states rather than intrinsically higher susceptibilities to adaptive convergent evolution, because morphological characters are no more prone to convergence than molecular characters once the number of states is controlled for.

Because the vast majority of molecular convergences are explainable by chance (Foote et al., 2015; Thomas and Hahn, 2015; Zou and Zhang, 2015a,b), the fact that average morphological characters have even smaller $\frac{Cv}{Dv}$ ratios than those of molecular characters of the same numbers of states suggests that most morphological convergences observed in the data analyzed are probably also attributable to chance. If convergence is owing to chance rather than lineage-specific selection, it is possible to identify and remove convergence-prone characters using species with reliable phylogenetic relationships and then infer the tree for species of uncertain relationships using the remaining characters. This approach would be

especially beneficial to phylogenetic inference that includes morphological data because of the relatively frequent convergence in such data. Zou and Zhang proposed a method to identify convergence-susceptible (morphological or molecular) characters and demonstrated that removing such characters improves phylogenetic accuracy (Zou and Zhang, 2016). Interestingly, applying this method to O'Leary et al.'s data alters the phylogenetic relationships among echolocating bats (Zou and Zhang, 2016).

5 Conclusions

Sequence convergence in any given gene is generally rare. However, when the entire genome is analyzed, hundreds of sites may show convergence. But because some neutral models predict even more convergence events than what has been observed, the vast majority of convergences observed in genome-wide analysis are attributable to chance. Nevertheless, this conclusion about sequence convergence at the genomic scale does not exclude the possibility of some adaptive events of sequence convergence. In fact, adaptive sequence convergence has been clearly demonstrated by statistical and experimental tests in a few genes. Experience suggests that genome-wide identification of sequence convergence, coupled with considerations of gene functions and relevant phenotypic effects, can provide candidates for adaptive convergence that should be followed up with experimental validation.

Appropriately modelling sequence evolution in the absence of positive selection is critical for a proper detection of adaptive convergence. This is a major methodological issue in current, and presumably future, literature on the subject. The processes of incomplete lineage sorting (Mendes et al., 2016) and introgression (Witt and Huerta-Sanchez, 2019) complicate the identification of genuine convergence events between closely related species (Lee and Coop, 2019).

Apart from potential indications of adaptation, convergence is a major source of phylogenetic noise. Comparative analyses of a large dataset of morphological and molecular characters used by systematists for inferring the mammalian phylogeny showed that morphological characters experienced more convergent evolution than molecular characters. Hence, molecular trees are expected to be more reliable than morphological trees with comparable data sizes. Interestingly, however, the reason behind the higher convergence of morphological than molecular characters is not that morphological characters are intrinsically more prone to convergence as a result of frequent positive selection. Instead, at least for the O'Leary et al. (2013) data, the reason is that morphological characters used by systematists tend to have fewer states than molecular characters, and the propensity for convergence is not higher for morphological than molecular characters once the number of states is controlled for. It has been shown than convergence-prone characters can be identified and removed to improve the accuracy of phylogenetic inference. This practice would be especially important for phylogenetic analysis involving morphological characters due to their higher probability of convergence. While the rapid accumulation of genome sequences will eventually dwarf the morphological data of any extant species, morphological data will remain useful in phylogenetic analysis that needs to contain fossils (see Chapter 5.1 [Pett and Heath 2020]), whose value to understanding evolution is indispensable. In this sense, better modeling of morphological convergence and development of methods for detecting convergence-prone traits will potentially improve the accuracy of phylogenetic reconstruction.

References

Arendt, J. and Reznick, D. (2008). Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends Ecol Evol*, 23(1):26–32.

Blackledge, T. A. and Gillespie, R. G. (2004). Convergent evolution of behavior in an adaptive radiation of hawaiian web-building spiders. *Proc Natl Acad Sci U S A*, 101(46):16228–33.

Castoe, T. A., de Koning, A. P. J., Kim, H. M., Gu, W. J., Noonan, B. P., Naylor, G., Jiang, Z. J., Parkinson, C. L., and Pollock, D. D. (2009). Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci U S A*, 106(22):8986–91.

Chikina, M., Robinson, J. D., and Clark, N. L. (2016). Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Mol Biol Evol*, 33(9):2182–92.

Christin, P. A., Salamin, N., Muasya, A. M., Roalson, E. H., Russier, F., and Besnard, G. (2008). Evolutionary switch and genetic convergence on rbcl following the evolution of c4 photosynthesis. *Mol Biol Evol*, 25(11):2361–8.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. J. Murray, London,.

Davalos, L. M., Cirranello, A. L., Geisler, J. H., and Simmons, N. B. (2012). Understanding phylogenetic incongruence: lessons from phyllostomid bats. *Biol. Rev. Camb. Philos. Soc.*, 87(4):991–1024.

Davalos, L. M., Velazco, P. M., Warsi, O. M., Smits, P. D., and Simmons, N. B. (2014). Integrating incomplete fossils by isolating conflicting signal in saturated and non-independent morphological characters. *Syst. Biol.*, 63(4):582–600.

Doolittle, R. F. (1994). Convergent evolution: the need to be explicit. *Trends Biochem Sci*, 19(1):15–8.

Feldman, C. R., Brodie, E. D., Brodie, E. D., and Pfrender, M. E. (2012). Constraint shapes convergence in tetrodotoxin-resistant sodium channels of snakes. *Proc Natl Acad Sci U S A*, 109(12):4556–61.

Fong, S. S., Joyce, A. R., and Palsson, B. O. (2005). Parallel adaptive evolution cultures of escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome Res*, 15(10):1365–72.

Foote, A. D., Liu, Y., Thomas, G. W. C., Vinar, T., Alfoldi, J., Deng, J. X., Dugan, S., van Elk, C. E., Hunter, M. E., Joshi, V., Khan, Z., Kovar, C., Lee, S. L., Lindblad-Toh, K., Mancia, A., Nielsen, R., Qin, X., Qu, J. X., Raney, B. J., Vijay, N., Wolf, J. B. W., Hahn, M. W., Muzny, D. M., Worley, K. C., Gilbert, M. T. P., and Gibbs, R. A. (2015). Convergent evolution of the genomes of marine mammals. *Nat. Genet.*, 47(3):272–5.

Gaubert, P., Wozencraft, W. C., Cordeiro-Estrela, P., and Veron, G. (2005). Mosaics of convergences and noise in morphological phylogenies: what's in a viverrid-like carnivoran? *Syst. Biol.*, 54(6):865–94.

Gerrard, E., Mutt, E., Nagata, T., Koyanagi, M., Flock, T., Lesca, E., Schertler, G. F. X., Terakita, A., Deupi, X., and Lucas, R. J. (2018). Convergent evolution of tertiary structure in rhodopsin visual proteins from vertebrates and box jellyfish. *Proc Natl Acad Sci U S A*, 115(24):6201–6.

Givnish, T. J. and Sytsma, K. J. (1997). Consistency, characters, and the likelihood of correct phylogenetic inference. *Mol. Phylogenet. Evol.*, 7(3):320–30.

Goldstein, R. A., Pollard, S. T., Shah, S. D., and Pollock, D. D. (2015). Non-adaptive amino acid convergence rates decrease over time. *Mol Biol Evol*, 32(6):1373–81.

Grenier, J. L. and Greenberg, R. (2005). A biogeographic pattern in sparrow bill morphology: parallel adaptation to tidal marshes. *Evolution*, 59(7):1588–95.

Jarvis, E. D., Mirarab, S., [...], Gilbert, M. T. P., and Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–31.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, 8(3):275–82.

Jost, M. C., Hillis, D. M., Lu, Y., Kyle, J. W., Fozzard, H. A., and Zakon, H. H. (2008). Toxin-resistant sodium channels: parallel adaptive evolution across a complete gene family. *Mol Biol Evol*, 25(6):1016–24.

Jousselin, E., Rasplus, J. Y., and Kjellberg, F. (2003). Convergence and coevolution in a mutualism: evidence from a molecular phylogeny of ficus. *Evolution*, 57(6):1255–69.

Langerhans, R. B., Knouft, J. H., and Losos, J. B. (2006). Shared and unique features of diversification in greater antillean anolis ecomorphs. *Evolution*, 60(2):362–9.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, 21(6):1095–109.

Lartillot, N. and Philippe, H. (2006). Computing bayes factors using thermodynamic integration. *Syst Biol*, 55(2):195–207.

Lee, K. M. and Coop, G. (2019). Population genomics perspectives on convergent adaptation. *Philos Trans R Soc Lond B Biol Sci*, 374(1777):20180236.

Legg, D. A., Sutton, M. D., and Edgecombe, G. D. (2013). Arthropod fossil data increase congruence of morphological and molecular phylogenies. *Nat. Commun.*, 4:2485.

Li, Y., Liu, Z., Shi, P., and Zhang, J. (2010). The hearing gene prestin unites echolocating bats and whales. *Curr Biol*, 20(2):R55–6.

Linnen, C. R., Poh, Y. P., Peterson, B. K., Barrett, R. D. H., Larson, J. G., Jensen, J. D., and Hoekstra, H. E. (2013). Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science*, 339(6125):1312–6.

Liu, Y., Cotton, J. A., Shen, B., Han, X., Rossiter, S. J., and Zhang, S. (2010). Convergent sequence evolution between echolocating bats and dolphins. *Curr Biol*, 20(2):R53–4.

Liu, Z., Li, S., Wang, W., Xu, D., Murphy, R. W., and Shi, P. (2011). Parallel evolution of kcnq4 in echolocating bats. *PLoS One*, 6(10):e26618.

Liu, Z., Qi, F. Y., Zhou, X., Ren, H. Q., and Shi, P. (2014). Parallel sites implicate functional convergence of the hearing gene prestin among echolocating mammals. *Mol Biol Evol*, 31(9):2415–24.

Manceau, M., Domingues, V. S., Linnen, C. R., Rosenblum, E. B., and Hoekstra, H. E. (2010). Convergence in pigmentation at multiple levels: mutations, genes and function. *Philos Trans R Soc Lond B Biol Sci*, 365(1552):2439–50.

Maruyama, M. and Parker, J. (2017). Deep-time convergence in rove beetle symbionts of army ants. *Curr Biol*, 27(6):920–6.

Melville, J., Harmon, L. J., and Losos, J. B. (2006). Intercontinental community convergence of ecology and morphology in desert lizards. *Proc Biol Sci*, 273(1586):557–63.

Mendes, F. K., Hahn, Y., and Hahn, M. W. (2016). Gene tree discordance can generate patterns of diminishing convergence over time. *Mol Biol Evol*, 33(12):3299–307.

Moore, J. and Willmer, P. (1997). Convergent evolution in invertebrates. *Biol Rev*, 72(1):1–60.

Natarajan, C., Hoffmann, F. G., Weber, R. E., Fago, A., Witt, C. C., and Storz, J. F. (2016). Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science*, 354(6310):336–9.

Nevo, E. (1979). Adaptive convergence and divergence of subterranean mammals. *Annu. Rev. Ecol. Evol. Syst.*, 10:269–308.

O'Leary, M. A., Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., Goldberg, S. L., Kraatz, B. P., Luo, Z. X., Meng, J., Ni, X. J., Novacek, M. J., Perini, F. A., Randall, Z. S., Rougier, G. W., Sargis, E. J., Silcox, M. T., Simmons, N. B., Spaulding, M., Velazco, P. M., Weksler, M., Wible, J. R., and Cirranello, A. L. (2013). The placental mammal ancestor and the post-k-pg radiation of placentals. *Science*, 339(6120):662–7.

Page, R. D. M. and Holmes, E. C. (1998). *Molecular evolution: a phylogenetic approach*. Blackwell Science.

Parker, J., Tsagkogeorga, G., Cotton, J. A., Liu, Y., Provero, P., Stupka, E., and Rossiter, S. J. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, 502(7470):228–31.

Perelman, P., Johnson, W. E., Roos, C., Seuanez, H. N., Horvath, J. E., Moreira, M. A., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M. P., Silva, A., O'Brien, S. J., and Pecon-Slattery, J. (2011). A molecular phylogeny of living primates. *PLoS Genet.*, 7(3):e1001342.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Projecto-Garcia, J., Natarajan, C., Moriyama, H., Weber, R. E., Fago, A., Cheviron, Z. A., Dudley, R., McGuire, J. A., Witt, C. C., and Storz, J. F. (2013). Repeated elevational transitions in hemoglobin function during the evolution of andean hummingbirds. *Proc Natl Acad Sci U S A*, 110(51):20669–74.

Protas, M. E., Hersey, C., Kochanek, D., Zhou, Y., Wilkens, H., Jeffery, W. R., Zon, L. I., Borowsky, R., and Tabin, C. J. (2006). Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nat Genet*, 38(1):107–11.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Robinson-Rechavi, M. (2020). Molecular evolution and gene function. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.2, pages 4.2:1–4.2:20. No commercial publisher | Authors open access book.

Rokas, A. and Carroll, S. B. (2008). Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol*, 25(9):1943–53.

Rosenblum, E. B., Rompler, H., Schoneberg, T., and Hoekstra, H. E. (2010). Molecular and functional basis of phenotypic convergence in white lizards at white sands. *Proc Natl Acad Sci U S A*, 107(5):2113–7.

Shen, Y. Y., Liu, J., Irwin, D. M., and Zhang, Y. P. (2010). Parallel and convergent evolution of the dim-light vision gene rh1 in bats (order: Chiroptera). *PLoS One*, 5(1):e8838.

Slot, J. C. and Rokas, A. (2010). Multiple gal pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc Natl Acad Sci U S A*, 107(22):10136–41.

Springer, M. S., Meredith, R. W., Teeling, E. C., and Murphy, W. J. (2013). Technical comment on "the placental mammal ancestor and the post-k-pg radiation of placentals". *Science*, 341(6146):613.

Stewart, C. B., Schilling, J. W., and Wilson, A. C. (1987). Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature*, 330(6146):401–4.

Thomas, G. W. and Hahn, M. W. (2015). Determining the null model for detecting adaptive convergence from genomic data: A case study using echolocating mammals. *Mol Biol Evol*, 32(5):1232–6.

Ujvari, B., Casewell, N. R., Sunagar, K., Arbuckle, K., Wuster, W., Lo, N., O'Meally, D., Beckmann, C., King, G. F., Deplazes, E., and Madsen, T. (2015). Widespread convergence in toxin resistance by predictable molecular evolution. *Proc Natl Acad Sci U S A*, 112(38):11911–6.

Vitali, D. G., Kaser, S., Kolb, A., Dimmer, K. S., Schneider, A., and Rapaport, D. (2018). Independent evolution of functionally exchangeable mitochondrial outer membrane import complexes. *Elife*, 7:e34488.

Wake, D. B., Wake, M. H., and Specht, C. D. (2011). Homoplasy: from detecting pattern to determining process and mechanism of evolution. *Science*, 331(6020):1032–5.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*, 18(5):691–9.

Wiens, J. J., Kuczynski, C. A., Townsend, T., Reeder, T. W., Mulcahy, D. G., and Sites, J. W., J. (2010). Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: Molecular data change the placement of fossil taxa. *Syst. Biol.*, 59(6):674–88.

Witt, K. E. and Huerta-Sanchez, E. (2019). Convergent evolution in human and domesticate adaptation to high-altitude environments. *Philos Trans R Soc Lond B Biol Sci*, 374(1777):20180235.

Wittkopp, P. J., Williams, B. L., Selegue, J. E., and Carroll, S. B. (2003). Drosophila pigmentation evolution: Divergent genotypes underlying convergent phenotypes. *Proc Natl Acad Sci U S A*, 100(4):1808–13.

Xu, S. H., He, Z. W., Guo, Z. X., Zhang, Z., Wyckoff, G. J., Greenberg, A., Wu, C. I., and Shi, S. H. (2017). Genome-wide convergence during evolution of mangroves from woody plants. *Mol Biol Evol*, 34(4):1008–15.

Zhang, J. (2003). Parallel functional changes in the digestive rnases of ruminants and colobines by divergent amino acid substitutions. *Mol Biol Evol*, 20(8):1310–7.

Zhang, J. (2006). Parallel adaptive origins of digestive rnases in asian and african leaf monkeys. *Nat Genet*, 38(7):819–23.

Zhang, J. and Kumar, S. (1997). Detection of convergent and parallel evolution at the amino acid sequence level. *Mol Biol Evol*, 14(5):527–36.

Zhang, J., Zhang, Y. P., and Rosenberg, H. F. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet*, 30(4):411–5.

Zhen, Y., Aardema, M. L., Medina, E. M., Schumer, M., and Andolfatto, P. (2012). Parallel molecular evolution in an herbivore community. *Science*, 337(6102):1634–7.

Zhou, X. M., Seim, I., and Gladyshev, V. N. (2015). Convergent evolution of marine mammals is associated with distinct substitutions in common genes. *Sci Rep*, 5:16550.

Zou, Z. and Zhang, J. (2015a). Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol Biol Evol*, 32(8):2085–96.

Zou, Z. and Zhang, J. (2015b). No genome-wide protein sequence convergence for echolocation. *Mol Biol Evol*, 32(5):1237–41.

Zou, Z. and Zhang, J. (2016). Morphological and molecular convergences in mammalian phylogenetics. *Nat Commun*, 7:12758.

Zou, Z. and Zhang, J. (2017). Gene tree discordance does not explain away the temporal decline of convergence in mammalian protein sequence evolution. *Mol Biol Evol*, 34(7):1682–8.

Zou, Z. and Zhang, J. (2019). Amino acid exchangeabilities vary across the tree of life. *Sci. Adv.*, 5(12):eaax3124.

Chapter 5.1 Inferring the Timescale of Phylogenetic Trees from Fossil Data

Walker Pett
Department of Ecology, Evolution, and Organismal Biology
Iowa State University, Ames, Iowa, 50011 USA
willpett@iastate.edu
https://orcid.org/0000-0003-3733-0815

Tracy A. Heath
Department of Ecology, Evolution, and Organismal Biology
Iowa State University, Ames, Iowa, 50011 USA
phylo@iastate.edu
https://orcid.org/0000-0002-0087-2541

Abstract

Time-stamped historical observations are required for scaling phylogenetic estimates to absolute time and, as a consequence, genomic data alone are not sufficient for dating the tree of life. The fossil record is the primary source of dated evidence of lineages over time and several statistical models for integrating paleontological and neontological data have been introduced. This chapter provides an overview of how fossil data are recovered from the rock record. We then describe two approaches to dating phylogenetic trees: (1) node dating where fossils are treated as calibrations for speciation times in an extant phylogeny and (2) the fossilized birth-death process as a mechanistic model that accounts for lineage diversification and fossil sampling. We conclude by discussing promising extensions of diversification models that can account for the structure of the fossil record and enable a more complete treatment of extinct and modern taxa in macroevolutionary analyses.

How to cite: Walker Pett and Tracy A. Heath (2020). Inferring the Timescale of Phylogenetic Trees from Fossil Data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 5.1, pp. 5.1:1–5.1:18. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Funding This work was supported by the National Science Foundation (USA) grants DEB-1556615, DEB-1556853, and DBI-1759909.

1 Introduction

Reconstructing the timescale of the tree of life is fundamental to understanding the pattern and process of species diversification. Inferring the topology of evolutionary relationships among species has been greatly facilitated by the advancements of the genomic era. These innovations include the accumulation of vast quantities of genomic character data, the development of high-dimensional statistical models of molecular evolution (Chapters 1.1 and 1.4 [Pupko and Mayrose 2020; Lartillot 2020a]), and increasingly robust computational tools (e.g. Chapters 1.3 and 1.5 [Kozlov and Stamatakis 2020; Lartillot 2020b]). However, estimating the timing of species divergences remains notoriously difficult, in part because good estimates can usually only be obtained by considering multiple sources of information simultaneously.

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 5.1; pp. 5.1:1–5.1:18

A book completely handled by researchers.
No publisher has been paid.

This is because phylogenetic character data—for example from molecular sequences or species morphology—can only tell us about the total amount of evolutionary change, or distance, among lineages. Evolutionary distances are traced out at a certain rate over a certain amount of time, and consequently information about the timing of species divergence is inherently confounded with information about the rate of evolutionary change (Chapter 3.2 [Boussau and Scornavacca 2020]). For this reason, independent estimates of rates and times cannot be obtained from phylogenetic character data alone. Thus, while the increasing sizes of molecular datasets may provide precise distance estimates, other sources of information about rates and times are ultimately required to improve the precision of divergence time estimates (dos Reis and Yang, 2013).

The fossil record provides the richest source of information about the absolute timing of lineage diversification in the tree of life. A fossil specimen is direct evidence for the existence of an ancient lineage and allows us to observe its age and preserved characteristics. These observations can then directly inform the phylogenetic placement of fossil specimens as well as the timing of their divergence from related species. By combining paleontological (fossil) and neontological (extant) data in a joint phylogenetic analysis, we can obtain independent estimates of molecular evolutionary rates and divergence times.

In order to include timing information from the fossil record in a phylogenetic analysis of molecular sequence data, a model is needed to describe the process of collecting and observing fossil data through time (Section 2). This model can range in complexity from a phenomenological description of the data (Section 3), to a richer mechanistic model of species diversification and fossil recovery (Section 4). By integrating these models of fossil occurrence time data with other types of data in a hierarchical Bayesian framework, the task of statistically disentangling rate and time is greatly simplified (Section 5).

2 Formalizing our Knowledge of the Fossil Record

Our understanding of the history of life on Earth begins with the geology of Earth's crust. Organisms bury evidence of their existence in the soils and sediments of their local environment, and over time these remains are compacted and preserved in rock layers. Eventually, geological or meteorological activity can expose these rock layers at the surface, where they can be observed and their ages determined, using radiometry or stratigraphic methods. When a fossil is discovered, its age can often only be determined as falling somewhere within the minimum and maximum extent of the rock layer containing it. If a fossil species spans multiple layers, its stratigraphic age range is determined from the ages of those layers. We will summarize the information contained in these fossil age observations using the symbol $\mathcal{F}$. Other observations not related to a fossil's age can also offer a wealth of phylogenetically informative characters, such as data from morphological and life history traits, biogeography, or even ancient DNA sequences (which have been recovered from specimens as old as 700,000 years; Orlando et al., 2013). These latter character observations are collectively denoted by the symbol $\mathcal{D}$, which may also include observations from extant species.

As mentioned already, evolutionary rates and times are inherently confounded as evolutionary distances. Inferences about divergence times from character data $\mathcal{D}$ alone will therefore be informed entirely by our *a priori* assumptions about the joint process of character evolution and species diversification. To get independent estimates of time, we must summarize our knowledge of the process of fossil preservation and collection generating not only the observed character data $\mathcal{D}$ but also the age data $\mathcal{F}$. Specifically, it will be convenient to formalize

our assumptions using a probabilistic mathematical model. In Bayesian parlance, the model is specified quantitatively as a joint *prior distribution* over the evolutionary parameters of interest, summarizing our knowledge of the model parameters prior to any data collection. Then, after collecting observations from the fossil record, we use Bayesian statistics to update our knowledge by estimating the *posterior distribution* of the model parameters.

2.1 Prior distribution on divergence times

We begin by specifying a prior distribution over our model parameters. In the context of divergence time estimation, our model can be thought of as broadly consisting of two sets of parameters. One set, labeled $\mathcal{T}$, includes those related to the diversification and species sampling process, such as the tree topology, divergence times, fossil sampling rates, etc. The other set, which we'll call θ, includes those related to the process of character evolution, such as rates of morphological evolution, or molecular substitution rates. We specify our prior distribution as the product of independent densities $f(\mathcal{T})$ and $f(\theta)$, such that

$$f(\mathcal{T}, \theta) = f(\mathcal{T})f(\theta). \tag{1}$$

The density $f(\mathcal{T})$ can be defined using a stochastic branching process like the Yule (Yule, 1924) or birth-death processes (Kendall, 1948). The density $f(\theta)$ can be defined in various ways (e.g. using a relaxed or other clock model, see Chapter 4.4 [Bromham 2020]) to describe the processes generating observed characters.

2.2 Posterior distribution on divergence times

To construct the posterior distribution over $\mathcal{T}$ and θ, we collect observations in the form of character data $\mathcal{D}$ and timing data from the fossil record $\mathcal{F}$. Then, using Bayes' theorem, the posterior distribution is proportional to the product of the prior distribution over $\mathcal{T}, \theta$ and the likelihood of $\mathcal{D}$ and $\mathcal{F}$

$$f(\mathcal{T}, \theta \mid \mathcal{D}, \mathcal{F}) \propto f(\mathcal{D} \mid \mathcal{T}, \theta) f(\mathcal{F} \mid \mathcal{T}) f(\mathcal{T}) f(\theta), \tag{2}$$

where the term $f(\mathcal{D} \mid \mathcal{T}, \theta)$ is the likelihood of the observed character data and the term $f(\mathcal{F} \mid \mathcal{T})$ is the likelihood of the observed fossil age data.

Importantly, we assume that the likelihood of the character data depends on $\mathcal{T}$ and θ, while the likelihood of the fossil age data $\mathcal{F}$ depends only on $\mathcal{T}$. The development of methods for calibrating trees to absolute time scales is primarily concerned with definitions of $f(\mathcal{T})$ and $f(\mathcal{F} \mid \mathcal{T})$. In the following sections, we will discuss two main approaches for defining these densities.

3 Node Calibration Densities

Historically, the most common statistical approach to inferring divergence times on a phylogenetic tree has been through the use of node calibrations. In this approach, information from the fossil record about the age of a particular clade is used to directly constrain the age of a node in the tree (typically the most-recent-common ancestor of the clade) during a phylogenetic analysis. This can be framed in a probabilistic approach by associating each calibrated node age with a probability density function, or node calibration density. In Bayesian inference, these densities are then used to compute the posterior distribution over $\mathcal{T}$, the tree topology and node ages. The way in which these node calibration densities are specified and applied has been the subject of a wide array of empirical and methodological studies (*e.g.*, Ho and Phillips, 2009; Warnock et al., 2012).

3.1 Conceptual formulation of node calibrations

Before exploring the ways in which node calibration densities can be applied in a phylogenetic analysis, we must first ask: *What, exactly, do node calibration densities represent?* There is considerable discussion of the representational and conceptual meaning of node calibration densities, but there are essentially two basic interpretations.

One interpretation formulates node calibrations directly as prior densities on the node ages (Yang and Rannala, 2005; Heled and Drummond, 2012). That is, the prior density $f(\mathcal{T})$ is constructed in such a way as to simultaneously account for uncertainty in the ages of both calibrated and uncalibrated nodes. Methods using this interpretation derive a conditional density on the uncalibrated node ages with fixed ages for the calibrated nodes, and then define the marginal prior on the calibrated node ages using a calibration density. It has been shown that for multiple node calibrations, this type of conditional prior leads to counterintuitive topologically inconsistent realized priors (Rannala, 2016), and may be computationally intractable (dos Reis, 2016).

For this reason, it is perhaps conceptually simpler to take the second point of view, which interprets node calibrations instead as representing the likelihood of the fossil data $f(\mathcal{F} \mid \mathcal{T})$. In this interpretation, the fossil likelihood is typically formulated as a product of marginal densities for each node in the tree for which there is relevant fossil data. In other words, for each calibrated node i with divergence time t_i, it is assumed that the likelihood of the fossil observations for that node is proportional to some density g_i. The full fossil likelihood is then a product of these node densities

$$f(\mathcal{F} \mid \mathcal{T}) \propto \prod_i g_i(t_i \mid \alpha),$$

where the parameterization α and functional form of the density g_i is specified by the researcher. In this approach, the prior density on the node ages $f(\mathcal{T})$ can then be specified simply using a familiar uncalibrated tree-generating process, such as the Yule or birth-death process.

Despite the relatively straightforward distinction between the two approaches, there is still considerable misunderstanding surrounding the interpretation of node calibration densities, and the two are commonly confused. In particular, a common misconception arises when the fossil data likelihood for a particular node is interpreted as a prior density on the age of that node. This interpretation as a prior is sometimes described as "incoherent" because it leads to the specification of two independent prior densities on the node ages, one from the node calibration g_i, and one from the tree prior $f(\mathcal{T})$ (Heled and Drummond, 2012; Heath et al., 2014).

Fortunately, the consequences of this misinterpretation are purely conceptual and should not have any quantitative impact on the resulting inferences. As an example, consider the distribution on the waiting time until recovery of the first fossil sample after the divergence of a clade. Assuming the waiting time is exponentially distributed, the density is the same whether we define the distribution with respect to the clade divergence time or the fossil age (Figure 1). Thus, whether we think of the density as likelihood or prior, both will lead to identical posterior distributions. Nevertheless, it is useful to follow the more principled likelihood interpretation as this will lead to more coherent and consistent application of node calibration methods generally.

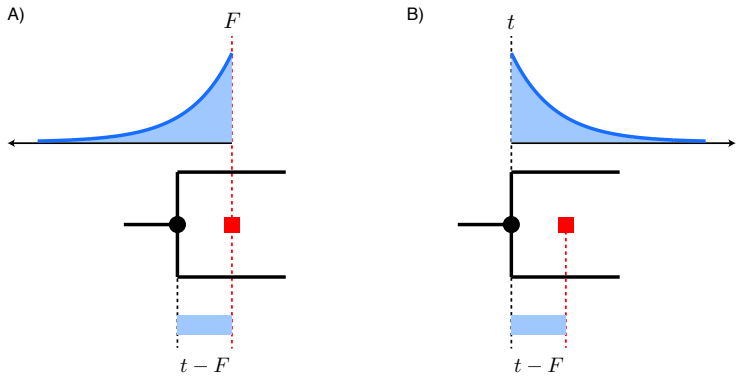


Figure 1 Alternative interpretations of a node calibration density as an (A) incoherent prior or a (B) likelihood. Given the age F of the first fossil specimen recovered after cladogenesis at time t, the waiting time $t - F$ is assumed to be exponentially distributed with rate λ. Both densities yield the same posterior distribution. A) The density is interpreted as a prior on the divergence time t. This results in the incoherent specification of two independent prior densities on t, coming from the node calibration as well as the tree prior. B) The density is interpreted as the likelihood fossil recovery age F.

3.2 Node calibrations with qualitative fossil data

In the simplest case, the form (*e.g.*, log-normal, exponential) and parameterization (*e.g.*, the mean, variance, and upper or lower bounds) of a node calibration density is chosen to qualitatively reflect the researcher's belief about the age of a clade based on their interpretation of the fossil data (Yang and Yoder, 2003; Yang and Rannala, 2005). In other words, the process of fossil preservation and observation is treated qualitatively, whereby the researcher's interpretation of these phenomena is implicit in the shape of $f(\mathcal{F} \mid \mathcal{T})$.

A variety of mathematical distributions have been proposed to represent these subjective calibrations with implicit fossil data, including normal, lognormal, exponential and uniform distributions (see Hedges and Kumar, 2004; Drummond et al., 2006; Donoghue and Benton, 2007; Ho, 2007; Ho and Phillips, 2009). These can be chosen such that the upper and/or lower bounds of the calibration density are "hard" or "soft" indicating whether there is assumed to be a non-zero probability of a fossil occurring outside the calibration bounds (Yang and Rannala, 2005; Sanders and Lee, 2007; Inoue et al., 2009). Soft bounds can be implemented for example by assuming the variance of the calibration density is such that 5% of the density falls beyond the maximum age constraint. While soft minimum bounds typically represent uncertainty in the age of the youngest calibration fossil (Benton and Donoghue, 2007), soft maximum bounds are typically justified either on the basis of models of diversification and preservation probability below the oldest known fossil in a clade (Foote et al., 1999; Tavaré et al., 2002), or using phylogenetic bracketing (Reisz and Müller, 2004; Müller and Reisz, 2005). Divergence time estimates can be extremely sensitive to the parameterization of the calibration density, but the impact on divergence time estimates of different prior densities is minimized when both minimum and maximum constraints are used (Warnock et al., 2012).

3.3 Node calibrations with quantitative fossil data

Other node calibration methods have been developed to make the representation of fossil data more quantitative and reproducible. For example, drawing on paleontological methods for estimating the stratigraphic ranges of fossil species (Strauss and Sadler, 1989), some node calibration methods make the explicit assumption that fossil recovery for a particular clade follows a constant-rate Poisson process through time, which implies that the ages of fossil specimens will be uniformly distributed over the clade's lifespan (Marshall, 2008; Dornburg et al., 2011; Wilkinson et al., 2011; Claramunt and Cracraft, 2015). Then, from the order statistics of a uniform distribution, it can be shown that the likelihood for the age of the oldest fossil F in a clade is equal to $f(F \mid t, n) = \frac{1}{t^n} n F^{n-1}$ (Strauss and Sadler, 1989), where t is the age of the clade and n is the number of fossil specimens. The calibration density $g(t \mid F, n)$ is proportional to the likelihood, and thus depends only on the number of fossil specimens

$$g(t \mid F, n) \propto f(F \mid t, n) \propto \frac{1}{t^n}, \quad t > F.$$

For example, Claramunt and Cracraft (2015) used this approach to calibrate the origin of modern birds.

Other methods construct node calibrations by modeling the process of fossil preservation as an exponential waiting time between clade divergence and fossil deposition (Wilkinson et al., 2011; Heath, 2012). Together, these methods take a step toward better formalizing the process of fossil data collection and interpretation, which ultimately makes their conclusions more testable and extensible.

Despite efforts to formalize the interpretation and characterization of node calibration densities, this approach to dating phylogenies still suffers from some limitations. Most notably, because calibration densities are only informed by the oldest node descended from a given calibrated node, these methods ignore much of the information present in the fossil record. Furthermore, fossil sampling times are observations of the underlying diversification process that gave rise to the phylogeny uniting the fossils and their extant relatives. In a statistical inference framework, these data can inform the parameters of the diversification model (*i.e.*, speciation and extinction), leading to more accurate and precise estimates.

3.4 Secondary calibrations

The posterior distribution summarizes our knowledge of the model parameters after taking some observations into account. If subsequent observations are made, the posterior can continue to be updated by considering it as a prior in relation to new data. This behavior can be leveraged to use posterior divergence time estimates from past studies as node calibrations in new analyses.

Specifically, consider the marginal posterior distribution $f(\mathcal{T} \mid \mathcal{F})$ obtained from a study using fossil data $\mathcal{F}$. This posterior is proportional to the fossil likelihood for $\mathcal{F}$ and the prior over $\mathcal{T}$

$$f(\mathcal{T} \mid \mathcal{F}) \propto f(\mathcal{F} \mid \mathcal{T}) f(\mathcal{T}). \tag{3}$$

Now imagine that new fossil observations $\mathcal{F}'$ are collected. We compute the joint posterior conditioned on both $\mathcal{F}$ and $\mathcal{F}'$ as

$$f(\mathcal{T} \mid \mathcal{F}, \mathcal{F}') \propto f(\mathcal{F}, \mathcal{F}' \mid \mathcal{T}) f(\mathcal{T}),$$

where $f(\mathcal{F}, \mathcal{F}' \mid \mathcal{T})$ is the joint likelihood of both sets of observations. If we assume that $\mathcal{F}$ and $\mathcal{F}'$ are sampled independently of each other, we may factor the likelihood such that

$$\begin{aligned} f(\mathcal{T} \mid \mathcal{F}, \mathcal{F}') &\propto f(\mathcal{F}' \mid \mathcal{T}) f(\mathcal{F} \mid \mathcal{T}) f(\mathcal{T}) \\ &\propto f(\mathcal{F}' \mid \mathcal{T}) f(\mathcal{T} \mid \mathcal{F}), \end{aligned}$$

where we have made use of Equation 3 to substitute the likelihood and prior terms involving $\mathcal{F}$ for the previously obtained marginal posterior $f(\mathcal{T} \mid \mathcal{F})$. The updated posterior can be estimated in a straightforward manner using MCMC and a previously obtained sample from $f(\mathcal{T} \mid \mathcal{F})$. In other words, we simply use the previously obtained posterior distribution as our new prior distribution on $\mathcal{T}$.

Importantly, in this approach we must consider the secondary node calibrations as part of the prior distribution over $\mathcal{T}$, and not as a fossil data likelihood. If we treated it as a likelihood and reweighed it according to an unconditioned prior, this would lead to the incoherent specification of two prior distribution terms for $\mathcal{T}$: one through the secondary calibration and one through our own specification of the prior. In order to avoid such incoherence in the prior, we must therefore either (1) specify the prior distribution as consisting entirely of the previously obtained sample, or (2) use an approach like those described at the beginning of Section 3.1 to condition the prior on the previous estimate (Yang and Rannala, 2005; Heled and Drummond, 2012). Despite this conceptual limitation of secondary calibrations, they are almost exclusively misapplied as fossil data likelihood terms, resulting in overly precise divergence time estimates (for review, see Schenk, 2016).

4 The Fossilized Birth-Death Process

As a result of the limitations of node calibration densities, it may be more satisfying to take an explicit, mechanistic approach by specifying a joint process $f(F, \mathcal{T})$ that simultaneously describes the generation of both the tree and the fossil data. Such an approach allows us to assume that fossil specimens are observations of lineages generated by the same diversification process that gave rise to the sampled living taxa. This allows for the inference of biologically meaningful parameters governing the diversification of both extant and extinct taxa—such as rates of speciation, extinction, and sampling—and leads to a more precise, quantitative representation of the process by which fossil specimens are sampled along lineages. Such a model was first described by Stadler (2010), who extended the birth-death process (Kendall, 1948; Nee et al., 1994; Gernhard, 2008; Stadler, 2009; Thompson, 1975) to account for lineages sampled back in time (see also Didier et al., 2012). By integrating fossil occurrence times into the branching model, this serially sampled birth-death process allows for estimation of macroevolutionary parameters under complex, mechanistic models of lineage diversification and fossil sampling.

4.1 Models for serially sampled data

Stadler (2010) introduced a serially sampled birth-death process that is well-suited to applications in macroevolution and in the study of infectious diseases (see Chapter 5.3 [Zhukova et al. 2020). When applied to macroevolutionary analyses of species-level data, this process requires samples from the fossil record and was thus coined the *fossilized birth-death process* (FBD) in Heath et al. (2014). Cladogenesis under a birth-death model begins with a single lineage that starts at time ϕ, this is the origin time of the process. Over the course of diversification, lineages speciate at rate λ and go extinct at rate μ. In the present (*i.e.*, $t = 0$),

every living lineage has a probability ρ of being sampled. In the absence of sampled fossils, all birth-death processes are governed by parameters λ, μ, and ρ. The FBD model includes a parameter for the rate of fossil recovery, denoted ψ, to account for observations from the fossil record. This parameter acts as a Poisson rate of sampling lineages over time.

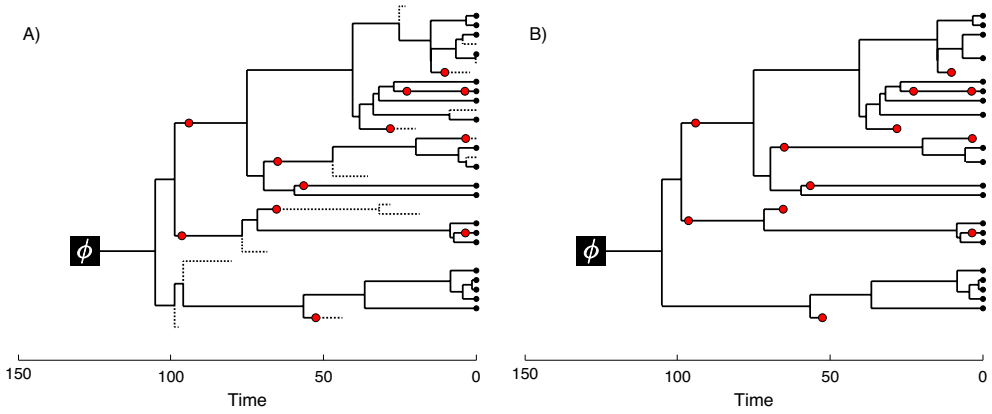


Figure 2 The fossilized birth-death process generates a (A) complete tree and a (B) sampled tree. (Figure modified from Figure 1 of Stadler, 2010)

The FBD model generates a *complete tree* and set of fossils Figure 2A. The *sampled tree* (also called the reconstructed tree) is the phylogeny after sampling, with all unobserved lineages pruned away Figure 2B. The probability density defined in Stadler (2010) allows us to compute the probability of any sampled tree while accounting for unobserved lineages in the complete tree. Figure 3 depicts the FBD process using a graphical model (for more on probabilistic graphical models for phylogenetics, see Höhna et al., 2014). This figure illustrates that the probability density of the sampled tree $\mathcal{T}$—which includes the tree topology, divergence times, and observed fossil occurrences $\mathcal{F}$—is dependent on the origin time ψ, speciation rate λ, extinction rate μ, fossilization rate ψ, and the extant species sampling probability ρ.

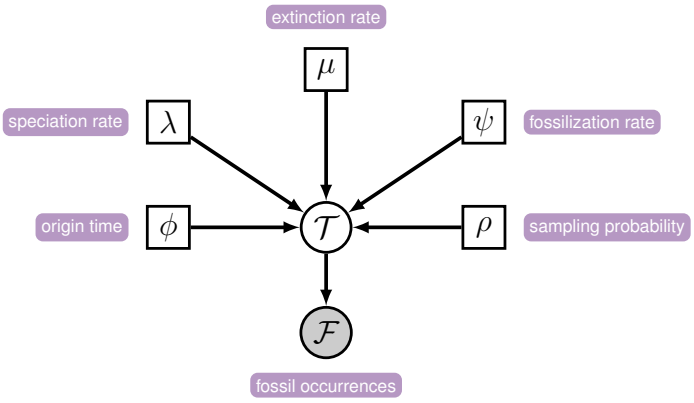


Figure 3 A graphical model depicting the structure of the fossilized birth-death process. The probability of the tree topology and divergence times $\mathcal{T}$ depend on the parameters of the FBD process: the time of origin ϕ, speciation rate λ, extinction rate μ, the rate of fossil recovery ψ, and the probability of sampling ρ. The observed fossil occurrences $\mathcal{F}$ are, in turn, dependent on $\mathcal{T}$ and the upstream parameters.

4.2 Sampled ancestors and the taxonomic assignment of fossil specimens

Under the FBD process, each fossil specimen is assumed to represent an independent sample from a continuously evolving lineage. Thus, there is a non-zero probability of obtaining a fossil sample that also has sampled descendant lineages Figure 2B. Indeed, Foote (1996) estimated that the probability of sampled ancestor-descendant pairs in the fossil record is non-negligible under a variety of cladogenetic models. Thus, it is important for diversification models to correctly account for sampled ancestors in order to accurately estimate speciation and extinction rates.

Under the FBD model, the proportion of fossil samples that also have sampled descendants is correlated with the probability of sampling extant lineages (ρ), the fossil sampling rate (ψ), and turnover ($r = \frac{\mu}{\lambda}$). We demonstrate this using simulations under the FBD model in Figure 4 for four different values of turnover and two different values of ρ, all over a range of values for ψ. There is a clear interplay between the parameters of the FBD model, which interact to yield different samples. Notably, even when extinction is relatively high ($r = 0.9$), extant sampling is low ($\rho = 0.1$), and the fossil recovery rate is low ($\psi = 0.01$), there is still a substantial proportion of fossils that also have sampled descendants (which may be fossil or extant samples), as seen in Figure 4.

Sampled ancestors present a curious taxonomic challenge. In the most straightforward formulation, no assumptions are made regarding the taxonomic assignment of sampled specimens, so all speciation events are assumed to correspond with branching events, and the speciation rate is therefore equal to the rate of cladogenesis. However, if two fossil samples are taken along a single evolving lineage, but are then assigned to different taxonomic species, then this implies an intervening anagenetic speciation event. Real speciation rates are therefore larger than the rate of cladogenesis, since some taxonomic species arise through anagenesis. In addition, the fossil record contains far more specimens than described taxa, and so many fossils are identified as belonging to the same taxonomic species. Consequently, in order to apply the FBD model to real fossil data, some assumptions must be made about the process of assigning fossil specimens to taxonomic species. Stadler et al. (2018) described a model for assigning fossil specimens to the same taxonomic species stratigraphic range.

4.3 Fossil placement under the fossilized birth-death process

Much like with node calibration approaches, divergence time estimates under the FBD may be sensitive to the phylogenetic placement of fossils. Unlike node calibrations, however, under the FBD model, all of the fossils—not just the oldest—that can be assigned to a node are valid observations. Nevertheless, it is important to consider the best practices outlined by Parham et al. (2012) when choosing fossils and justifying their placement in the phylogeny.

In many cases, quantitative character data have not been coded or are otherwise not available for a particular fossil taxon. In these instances, qualitative information about the topological placement of a fossil may be derived from its observed occurrence time and/or the taxonomic literature (Heath et al., 2014). For many fossils, it may be ambiguous whether the fossil lineage falls on the stem lineage or within the crown of an extant clade (Benton and Donoghue, 2007). In many cases, it is only possible to define the fossil and its relatives as a monophyletic "total group", within which there are many possible placements of the fossil as a crown or stem fossil. Bayesian inference methods using the FBD process can account for this uncertainty by integrating over the different possible fossil placements using Markov chain Monte Carlo (MCMC).

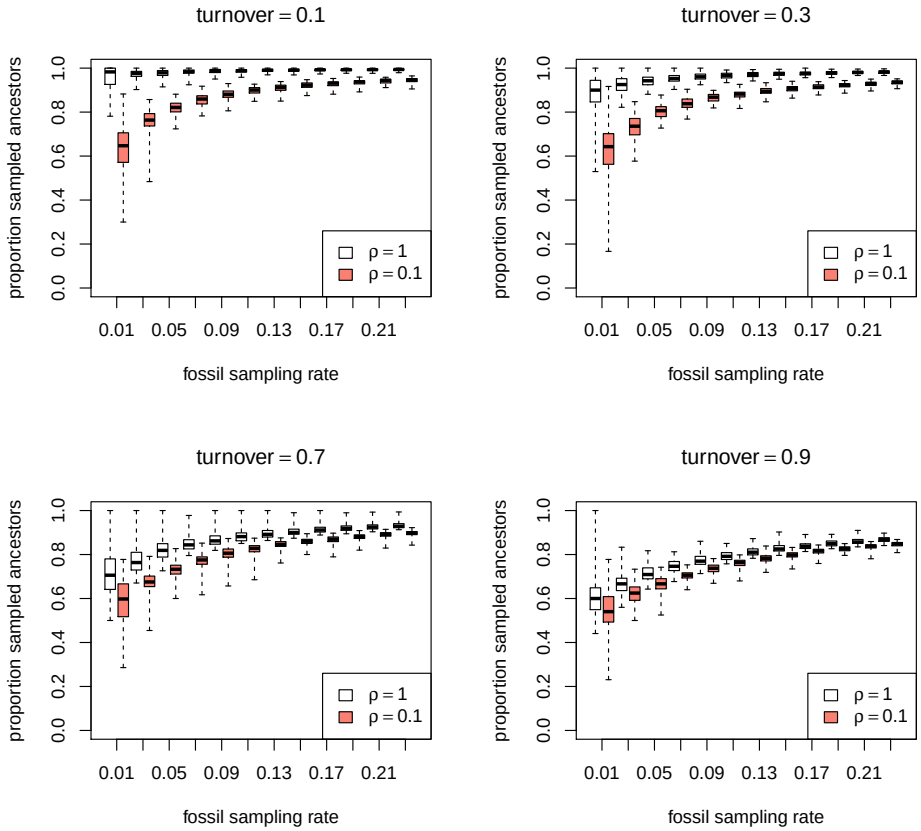


Figure 4 The proportion of sampled ancestors simulated under different FBD parameters. Here, *turnover* (r) is defined as $r = \frac{\mu}{\lambda}$, where λ and μ denote *speciation* and *extinction*, respectively. For each of the four values for turnover, we show the proportion of sampled ancestors for 100 simulated replicates as we varied the fossil sampling rate (ψ) and for two different values of the probability of sampling extant taxa (ρ).

When character data are available for fossil taxa, an integrative modeling approach is needed to combine observations from both extant and extinct species. The model and methods described in Stadler (2010), Ronquist et al. (2012), Zhang et al. (2016), and others provide a framework for using the FBD model in more fully integrative Bayesian analysis of fossil and extant samples (see section 5).

4.3.1 Empirical studies applying the fossilized birth-death model

Analysis under the FBD process enables researchers to use more of the data from the fossil record, which, in turn, can lead to more robust estimates and a more comprehensive understanding of lineage diversification. Using simulated trees and data, Heath et al. (2014) demonstrated that when using fossil occurrences to date extant phylogenies under the FBD model, node age estimates are more accurate than conventional calibration density approaches. Importantly, this study also showed that the precision of FBD node age estimates increases as the number of fossil occurrences increases, providing a better representation of statistical uncertainty in these parameters. Didier et al. (2017) also developed a maximum likelihood

approach to estimate parameters of the diversification model when fossil occurrences are observed. Their analyses of simulated datasets demonstrate that estimates of speciation and extinction rates are more accurate when fossil ages are included compared to estimates based on trees of extant taxa with the node ages fixed to their true values.

The FBD model has gained traction in both neontological and paleontological studies because its assumptions are more justifiable than node-calibration density approaches (as described in Section 3). As a result, empirical studies are emerging that provide new insights into the macroevolution of numerous clades in the tree of life. For example, the FBD process has been used for the calibration of extant phylogenies of royal ferns (Grimm et al., 2014), tetraodontiform fishes (Arcila et al., 2015), and pines (Saladin et al., 2017); in combined evidence analyses of hymenopterans (Zhang et al., 2016), lemurs (Herrera and Dávalos, 2016), myriapods (Fernández et al., 2016), sloths (Slater et al., 2016), penguins (Gavryushkina et al., 2017), baleen whales (Slater et al., 2017), and sponges (Schuster et al., 2018); or to study extinct clades using morphological characters and occurrence times of theropods (Bapst et al., 2016), and crinoids (Wright et al., 2017).

5 Integrative Hierarchical Models for Calibrating Time Trees

One important advantage of the fossilized birth-death modeling approach is that it makes our evolutionary analysis more integrative. By directly modeling fossil sampling jointly with cladogenesis and extinction, we can connect information from disparate evolutionary processes and synthesize all of it in a single *hierarchical* analysis. In a hierarchical model, relationships among collections of model parameters are structured in a directional, tree-like manner, such that information from a number of empirical observations of different datatypes can be considered jointly through their shared dependence on a smaller number of upstream model parameters. For example, the FBD model allows us to unify information from both the fossil record and the molecular record, by connecting models of fossil sampling and molecular evolution indirectly through a time tree model of speciation and extinction (as shown in the hierarchical model in Figure 5). In other words, the FBD model provides the foundation on which to construct much larger and more elaborate probabilistic models that link a wide range of information sources in a Bayesian hierarchical inference framework.

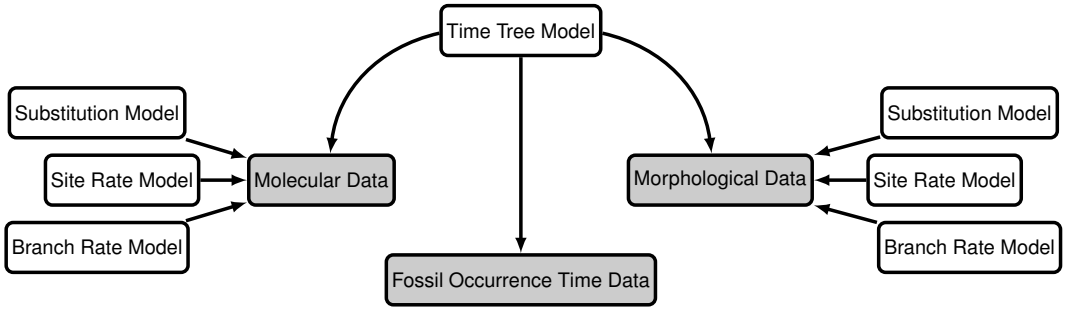


Figure 5 An example of a hierarchical model using molecular, morphological and fossil occurrence time data. Nodes represent collections of random variables and model parameters. Shaded nodes represent variables that are associated with empirical observations.

5.1 Integrating molecules and morphology

One immediate consequence of including fossil and extant taxa in a joint hierarchical analysis is that character data from both extinct and extant species can be combined. In particular, morphological data from fossil specimens can be directly combined with data from extant species. These *combined-evidence* (also called "total-evidence" or "tip-dating") approaches allow for inference of fossil species relationships with extant taxa (Ronquist et al., 2012). Thus, this integrative statistical approach leads to more reproducible placement of fossil species using explicit methodology, instead of reliance on previously published studies or taxonomy. The placement of calibration fossils using morphological data also results in more robust estimates of divergence times and diversification dynamics among extant species (Heath et al., 2014; Gavryushkina et al., 2014). Furthermore, any uncertainty in the placement of fossils can be accommodated by implementing this approach in a Bayesian framework, resulting in parameter estimates that better reflect the information in the data.

The first combined-evidence analyses described in the literature either did not use an appropriate diversification model (Pyron, 2011, used a pure-birth model that did not allow extinction or fossil sampling) or used a generic, phenomenological model (Ronquist et al., 2012, assumed a uniform model for node ages, fossil sampling times, and topologies). Nevertheless, these studies outlined a framework for a more fully integrative approach to inferring dated phylogenies of living and fossil taxa. With the introduction of new MCMC proposal mechanisms for FBD trees (Heath et al., 2014; Gavryushkina et al., 2014; Zhang et al., 2016), combined-evidence analyses can now include much more appropriate and realistic models of lineage diversification and sampling. Using a combined-evidence approach, the fossilized birth-death process has been applied to combined datasets from Hymenoptera (Zhang et al., 2016), penguins (Gavryushkina et al., 2017), mammals (Upham et al., 2019) and squamates (Pyron, 2016), to name a few. Luo et al. (2020) used simulations to explore the accuracy of combined-evidence dating approaches under the fossilized birth death process, and found that including morphological data led to more accurate estimation of divergence times, but that most dating information was contained in fossil occurrence age data.

5.2 Sampling bias

As our integrative modeling approach grows in complexity, and encompasses more and different types of data, it becomes extremely important to carefully consider the sampling strategies used to acquire those data, and ask whether any sampling biases may be impacting the analysis. If so, we must look for ways of explicitly accounting for the data sampling strategy when specifying the model. Here we discuss several important sources of sampling biases as they pertain to divergence time estimation and the fossilized birth-death process.

5.2.1 Taxonomic - diversified sampling

One important type of sampling bias may arise as a result of a commonly employed diversified taxon sampling strategy. In this approach, only representatives from major lineages are included in an analysis in an attempt to achieve broad taxonomic representation while at the same time minimizing the size of the dataset and redundancy in the analysis (*e.g.*, Jarvis et al., 2014, sampled a single representative from every avian order to estimate the relationships and divergence times of birds). This type of sampling will induce longer terminal branches than under a random sampling scheme (Heath, 2008). Höhna et al. (2011) described an approach using a birth-death process that accounts for diversified sampling by assuming that lineages are sampled such that the total phylogenetic diversity (tree length) is maximized.

Analytic expressions are available for the probability density of a birth-death tree under this type of diversified sampling (Höhna et al., 2011). Zhang et al. (2016) applied this approach to the FBD model by assuming that exactly one representative extant species per clade descending from some cutoff time is selected. Accounting for diversified sampling seems to be important in some empirical datasets, resulting in much younger age estimates for some clades (Vea and Grimaldi, 2016; Ronquist et al., 2016).

5.2.2 Macroevolutionary - conditioning on survival

Some clades are never sampled simply because they did not survive long enough to be observed in the fossil record. This leads to a systematic undersampling of clades undergoing relatively high rates of extinction, which can lead to a bias in the estimation of background speciation and extinction rates. This effect can be accounted for by explicitly computing the conditional probability of a clade, given that at least one sample was recovered. Guindon (2018) devised an MCMC algorithm for computing this conditional probability under the fossilized birth-death process. Usually, however, the background macroevolutionary regime of speciation and extinction is not of interest when conducting inference on a single clade, and indeed may be essentially impossible to estimate with any accuracy (see for example the Lartillot, 2014, blog post). Therefore, in most cases it is probably unnecessary to condition on survival.

5.2.3 Stratigraphic - fossil sampling through time

It is well known that the fossil record is incomplete and unevenly sampled, and a wide range of factors impact what organisms are preserved and how. Obviously, organisms like mammals and snails with hard parts like a skeleton or shell will be better preserved in the fossil record, and will therefore be systematically oversampled. Similarly, fossils deposited in the more recent past will also be better preserved, and are therefore more likely to be recovered. Many paleontological studies have attempted to account for these biases when reconstructing estimates of species diversity in ancient clades (Sepkoski et al., 1981; Foote and Sepkoski, 1999; Raup, 1972, 1976). These studies have often relied on estimates obtained by combining sampling effort or proxy data with fossil abundance data in a multivariate model, a method known as "residual diversity estimation" (Smith and McGowan, 2007; Sakamoto et al., 2017).

While no studies have applied a similar approach jointly with the fossilized birth-death process, the mathematics enabling the estimation of time-heterogeneous fossil sampling and diversification have been described for the FBD (Gavryushkina et al., 2014). Thus, implementing a fossilized birth-death variation of the residual diversity estimation approach is feasible within a hierarchical modeling framework, and will be a valuable goal of future studies.

6 Prospectus

As the field of statistical phylogenetics has matured in the genomic era, vast quantities of molecular data have become widely available for studying a variety of extant (and some recently extinct) species. This has made it possible to obtain very good estimates of the evolutionary relationships among many clades whose relationships were previously unknown due to a lack of phylogenetically informative characters coming from other sources, particularly species with no representation in the fossil record. Ultimately, however, genomic data alone are insufficient for resolving the absolute ages of species divergences. Thus,

although advancements in sequencing technologies have yielded data that provide great evolutionary insights, reconstruction of the macroevolutionary timeline is limited to classical—and often laborious—methods in paleontology for collecting and dating fossil specimens. Consequently, technological innovations in collecting, organizing, and curating paleontological data will be critical if we are to make major progress in elucidating the absolute divergence times of the tree of life. In clades with poor fossil records, the development of models that account for additional sources of dating information, such as biogeographic patterns (Landis, 2017), or patterns in the conservation of horizontal gene transfer events (Davín et al., 2018) will be crucial for inferring divergence times. However, the timescale of evolutionary events in some clades without good representation in the fossil or other type of historic record may never be fully understood.

7 Acknowledgements

We wish to thank the editors F. Delsuc, N. Galtier, and C. Scornavacca for the opportunity to contribute our paper to *Phylogenetics in the Genomic Era*. Our colleagues J. Barido-Sottani, J. Buckner, W. Dismukes, J. Justison, K. Quinteros, J. Satler, and D. Żyła provided helpful comments that greatly improved this chapter.

References

Arcila, D., Pyron, R. A., Tyler, J. C., Ortí, G., and Betancur-R, R. (2015). An evaluation of fossil tip-dating versus node-age calibrations in tetraodontiform fishes (Teleostei: Percomorphaceae). *Molecular Phylogenetics and Evolution*, 82:131–145.

Bapst, D., Wright, A., Matzke, N., and Lloyd, G. (2016). Topology, divergence dates, and macroevolutionary inferences vary between different tip-dating approaches applied to fossil theropods (Dinosauria). *Biology Letters*, 12(7):20160237.

Benton, M. J. and Donoghue, P. C. (2007). Paleontological evidence to date the tree of life. *Molecular Biology and Evolution*, 24(1):26–53.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Bromham, L. (2020). Substitution rate analysis and molecular evolution. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.4, pages 4.4:1–4.4:21. No commercial publisher | Authors open access book.

Claramunt, S. and Cracraft, J. (2015). A new time tree reveals Earth history's imprint on the evolution of modern birds. *Science Advances*, 1(11):e1501005.

Davín, A. A., Tannier, E., Williams, T. A., Boussau, B., Daubin, V., and Szöllősi, G. J. (2018). Gene transfers can date the tree of life. *Nature Ecology & Evolution*, 2(5):904.

Didier, G., Fau, M., and Laurin, M. (2017). Likelihood of tree topologies with fossils and diversification rate estimation. *Systematic Biology*, 66(6):964–987.

Didier, G., Royer-Carenzi, M., and Laurin, M. (2012). The reconstructed evolutionary process with the fossil record. *Journal of Theoretical Biology*, 315:26–37.

Donoghue, P. C. and Benton, M. J. (2007). Rocks and clocks: calibrating the tree of life using fossils and molecules. *Trends in Ecology & Evolution*, 22(8):424–431.

Dornburg, A., Beaulieu, J. M., Oliver, J. C., and Near, T. J. (2011). Integrating fossil preservation biases in the selection of calibrations for molecular divergence time estimation. *Systematic Biology*, 60(4):519–527.

dos Reis, M. (2016). Notes on the birth–death prior with fossil calibrations for Bayesian estimation of species divergence times. *Philosophical Transactions of the Royal Society B*, 371(1699):20150128.

dos Reis, M. and Yang, Z. (2013). The unbearable uncertainty of Bayesian divergence time estimation. *Journal of Systematics and Evolution*, 51(1):30–43.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88.

Fernández, R., Edgecombe, G. D., and Giribet, G. (2016). Exploring phylogenetic relationships within Myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. *Systematic Biology*, 65(5):871–889.

Foote, M. (1996). On the probability of ancestors in the fossil record. *Paleobiology*, 22:141–151.

Foote, M., Hunter, J. P., Janis, C. M., and Sepkoski, J. J. (1999). Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. *Science*, 283(5406):1310–1314.

Foote, M. and Sepkoski, J. J. (1999). Absolute measures of the completeness of the fossil record. *Nature*, 398(6726):415.

Gavryushkina, A., Heath, T. A., Ksepka, D. T., Stadler, T., Welch, D., and Drummond, A. J. (2017). Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic Biology*, 66(1):57–73.

Gavryushkina, A., Welch, D., Stadler, T., and Drummond, A. J. (2014). Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology*, 10(12):e1003919.

Gernhard, T. (2008). The conditioned reconstructed process. *Journal of Theoretical Biology*, 253:769–778.

Grimm, G. W., Kapli, P., Bomfleur, B., McLoughlin, S., and Renner, S. S. (2014). Using more than the oldest fossils: dating Osmundaceae with three Bayesian clock approaches. *Systematic Biology*, 64(3):396–405.

Guindon, S. (2018). Accounting for calibration uncertainty: Bayesian molecular dating as a "doubly intractable" problem. *Systematic Biology*, 67(4):651–661.

Heath, T. A. (2008). *Understanding the importance of taxonomic sampling for large-scale phylogenetic analyses by simulating evolutionary processes under complex models*. PhD thesis, University of Texas at Austin, https://repositories.lib.utexas.edu/handle/2152/18347.

Heath, T. A. (2012). A hierarchical Bayesian model for calibrating estimates of species divergence times. *Systematic Biology*, 61(5):793–809.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences, USA*, 111(29):E2957–E2966.

Hedges, S. B. and Kumar, S. (2004). Precision of molecular time estimates. *Trends in Genetics*, 20(5):242–247.

Heled, J. and Drummond, A. J. (2012). Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology*, 61(1):138–149.

Herrera, J. P. and Dávalos, L. M. (2016). Phylogeny and divergence times of lemurs inferred with recent and ancient fossils in the tree. *Systematic Biology*, 65(5):772–791.

Ho, S. Y. (2007). Calibrating molecular estimates of substitution rates and divergence times in birds. *Journal of Avian Biology*, 38(4):409–414.

Ho, S. Y. W. and Phillips, M. J. (2009). Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology*, 58:367–380.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., and Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic Biology*, 63(5):753–771.

Höhna, S., Stadler, T., Ronquist, F., and Britton, T. (2011). Inferring speciation and extinction rates under different sampling schemes. *Molecular Biology and Evolution*, 28(9):2577–2589.

Inoue, J., Donoghue, P. C., and Yang, Z. (2009). The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Systematic Biology*, 59(1):74–89.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y., Faircloth, B. C., Nabholz, B., Howard, J. T., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.

Kendall, D. G. (1948). On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, 19(1):1–15.

Kozlov, A. M. and Stamatakis, A. (2020). Using raxml-ng in practice. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.3, pages 1.3:1–1.3:25. No commercial publisher | Authors open access book.

Landis, M. J. (2017). Biogeographic dating of speciation times using paleogeographically informed processes. *Systematic Biology*, 66(2):128–144.

Lartillot, N. (2014). Should we condition on non-extinction? Blog: The Bayesian Kitchen (http://bayesiancook.blogspot.com/2014/11/should-we-condition-on-non-extinction.html).

Lartillot, N. (2020a). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. (2020b). Phylobayes: Bayesian phylogenetics using site-heterogeneous models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.5, pages 1.5:1–1.5:16. No commercial publisher | Authors open access book.

Luo, A., Duchêne, D. A., Zhang, C., Zhu, C.-D., and Ho, S. Y. W. (2020). A simulation-based evaluation of tip-dating under the fossilized birth–death process. *Systematic Biology*, 69(2):325–344.

Marshall, C. R. (2008). A simple method for bracketing absolute divergence times on molecular phylogenies using multiple fossil calibration points. *The American Naturalist*, 171(6):726–742.

Müller, J. and Reisz, R. R. (2005). Four well-constrained calibration points from the vertebrate fossil record for molecular clock estimates. *BioEssays*, 27(10):1069–1075.

Nee, S., May, R. M., and Harvey, P. H. (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society B*, 344:305–311.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, 499(7456):74.

Parham, J. F., Donoghue, P. C. J., Bell, C. J., Calway, T. D., Head, J. J., Holroyd, P. A., Inoue, J. G., Irmis, R. B., Joyce, W. G., Ksepka, D. T., Patané, J. S. L., Smith, N. D., Tarver, J. E., van Tuinen, M., Yang, Z., Angielczyk, K. D., Greenwood, J. M., Hipsley, C. A., Jacobs, L., Makovicky, P. J., Müller, J., Smith, K. T., Theodor, J. M., Warnock, R. C. M., and Benton, M. J. (2012). Best practices for justifying fossil calibrations. *Systematic Biology*, 61(2):346–359.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Pyron, R. A. (2011). Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Systematic Biology*, 60:466–481.

Pyron, R. A. (2016). Novel approaches for phylogenetic inference from morphological data and total-evidence dating in squamate reptiles (lizards, snakes, and amphisbaenians). *Systematic Biology*, 66(1):38–56.

Rannala, B. (2016). Conceptual issues in Bayesian divergence time estimation. *Philosophical Transactions of the Royal Society B*, 371(1699):20150134.

Raup, D. M. (1972). Taxonomic diversity during the Phanerozoic. *Science*, 177(4054):1065–1071.

Raup, D. M. (1976). Species diversity in the Phanerozoic: an interpretation. *Paleobiology*, 2(4):289–297.

Reisz, R. R. and Müller, J. (2004). Molecular timescales and the fossil record: a paleontological perspective. *Trends in Genetics*, 20(5):237–241.

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., and Rasnitsyn, A. P. (2012). A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology*, 61(6):973–999.

Ronquist, F., Lartillot, N., and Phillips, M. J. (2016). Closing the gap between rocks and clocks using total-evidence dating. *Philosophical Transactions of the Royal Society B*, 371(1699):20150136.

Sakamoto, M., Venditti, C., and Benton, M. J. (2017). 'Residual diversity estimates' do not correct for sampling bias in palaeodiversity data. *Methods in Ecology and Evolution*, 8(4):453–459.

Saladin, B., Leslie, A. B., Wüest, R. O., Litsios, G., Conti, E., Salamin, N., and Zimmermann, N. E. (2017). Fossils matter: improved estimates of divergence times in *Pinus* reveal older diversification. *BMC Evolutionary Biology*, 17(1):95.

Sanders, K. L. and Lee, M. S. (2007). Evaluating molecular clock calibrations using Bayesian analyses with soft and hard bounds. *Biology Letters*, 3(3):275–279.

Schenk, J. J. (2016). Consequences of secondary calibrations on divergence time estimates. *PLoS One*, 11(1):e0148228.

Schuster, A., Vargas, S., Knapp, I. S., Pomponi, S. A., Toonen, R. J., Erpenbeck, D., and Wörheide, G. (2018). Divergence times in demosponges (Porifera): first insights from new mitogenomes and the inclusion of fossils in a birth-death clock model. *BMC Evolutionary Biology*, 18(1):114.

Sepkoski, J. J., Bambach, R. K., Raup, D. M., and Valentine, J. W. (1981). Phanerozoic marine diversity and the fossil record. *Nature*, 293(5832):435.

Slater, G. J., Cui, P., Forasiepi, A. M., Lenz, D., Tsangaras, K., Voirin, B., de Moraes-Barros, N., MacPhee, R. D., and Greenwood, A. D. (2016). Evolutionary relationships among extinct and extant sloths: the evidence of mitogenomes and retroviruses. *Genome Biology and Evolution*, 8(3):607–621.

Slater, G. J., Goldbogen, J. A., and Pyenson, N. D. (2017). Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. *Royal Society B: Biological Sciences*, 284(1855):20170546.

Smith, A. B. and McGowan, A. J. (2007). The shape of the Phanerozoic marine palaeodiversity curve: how much can be predicted from the sedimentary rock record of Western Europe? *Palaeontology*, 50(4):765–774.

Stadler, T. (2009). On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58–66.

Stadler, T. (2010). Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*, 267(3):396–404.

Stadler, T., Gavryushkina, A., Warnock, R. C., Drummond, A. J., and Heath, T. A. (2018). The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. *Journal of Theoretical Biology*, 447:41–55.

Strauss, D. and Sadler, P. M. (1989). Classical confidence intervals and Bayesian probability estimates for ends of local taxon ranges. *Mathematical Geology*, 21(4):411–427.

Tavaré, S., Marshall, C. R., Will, O., Soligo, C., and Martin, R. D. (2002). Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature*, 416(6882):726.

Thompson, E. A. (1975). *Human Evolutionary Trees*. Cambridge University Press, Cambridge, UK.

Upham, N. S., Esselstyn, J. A., and Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biology*, 17(12):e3000494.

Vea, I. M. and Grimaldi, D. A. (2016). Putting scales into evolutionary time: the divergence of major scale insect lineages (Hemiptera) predates the radiation of modern angiosperm hosts. *Scientific Reports*, 6:23487.

Warnock, R. C., Yang, Z., and Donoghue, P. C. (2012). Exploring uncertainty in the calibration of the molecular clock. *Biology Letters*, 8(1):156–159.

Wilkinson, R. D., Steiper, M. E., Soligo, C., Martin, R. D., Yang, Z., and Tavaré, S. (2011). Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Systematic Biology*, 60:16–31.

Wright, D. F., Zamora, S., and Rahman, I. A. (2017). Bayesian estimation of fossil phylogenies and the evolution of early to middle Paleozoic crinoids (Echinodermata). *Journal of Paleontology*, 91(4):799–814.

Yang, Z. and Rannala, B. (2005). Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution*, 23(1):212–226.

Yang, Z. and Yoder, A. D. (2003). Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic Biology*, 52(5):705–716.

Yule, G. U. (1924). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Wills, F. R. S. *Philosophical Transactions of the Royal Society of London, Biology*, 213:21–87.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. (2016). Total-evidence dating under the fossilized birth-death process. *Systematic Biology*, 65(2):228–249.

Zhukova, A., Gascuel, O., Duchêne, S., Ayres, D. L., Lemey, P., and Baele, G. (2020). Efficiently analysing large viral data sets in computational phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.3, pages 5.3:1–5.3:43. No commercial publisher | Authors open access book.

Chapter 5.2 Estimating a Time-calibrated Phylogeny of Fossil and Extant Taxa using RevBayes

Joëlle Barido-Sottani
Department of Ecology, Evolution, & Organismal Biology
Iowa State University
Ames, IA 50011 USA
joellebs@iastate.edu
https://orcid.org/0000-0002-5220-5468

Joshua A. Justison
Department of Ecology, Evolution, & Organismal Biology
Iowa State University
Ames, IA 50011 USA
justison@iastate.edu
https://orcid.org/0000-0002-0233-4413

April M. Wright
Department of Biological Sciences
Southeastern Louisiana University
Hammond, LA 70402 USA
april.wright@selu.edu
https://orcid.org/0000-0003-4692-3225

Rachel C. M. Warnock
Department of Biosystems Science & Engineering
Eidgenössische Technische Hochschule Zürich
Swiss Institute of Bioinformatics (SIB)
4058 Basel, Switzerland
rachel.warnock@bsse.ethz.ch
https://orcid.org/0000-0002-9151-4642

Walker Pett
Department of Ecology, Evolution, & Organismal Biology
Iowa State University
Ames, IA 50011 USA
willpett@iastate.edu
https://orcid.org/0000-0003-3733-0815

Tracy A. Heath
Department of Ecology, Evolution, & Organismal Biology
Iowa State University
Ames, IA 50011 USA
phylo@iastate.edu
https://orcid.org/0000-0002-0087-2541

Abstract

The fossil record is the primary source of time-stamped information useful for dating phylogenetic trees; and many statistical approaches are available for integrating data from fossil and living species. In this tutorial, we demonstrate how to perform joint inference of divergence times and phylogenetic relationships of fossil and extant taxa from morphological data using the program RevBayes. RevBayes (`http://revbayes.com`) is a flexible and powerful tool for Bayesian

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 5.2; pp. 5.2:1–5.2:23

A book completely handled by researchers.
No publisher has been paid.

phylogenetic inference. Statistical models in RevBayes are built using probabilistic graphical models and described via an interpreted programming language. As a result, RevBayes offers a wide range of statistical models—ranging from very simple models with few parameters to hierarchical models describing complex biological processes—that are useful in many biological applications. The exercise described here provides instructions on how to construct a phylogenetic model combining the fossilized birth-death process and models describing the generation of morphological data, which is then used to execute an analysis that unites modern and extinct taxa in a dated phylogenetic tree. The content and associated files for this tutorial are kept up-to-date at: http://revbayes.com/tutorials/fbd_simple.

How to cite: Joëlle Barido-Sottani, Joshua A. Justison, April M. Wright, Rachel C. M. Warnock, Walker Pett, and Tracy A. Heath (2020). Estimating a time-calibrated phylogeny of fossil and extant taxa using RevBayes. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 5.2, pp. 5.2:2–5.2:23. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Funding This work was supported by National Science Foundation (USA) grants DEB-1556615, DEB-1556853, and DBI-1759909 (JBS, JAJ, WP, and TAH); and an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P2O GM103424-18 (AMW).

1 Introduction to RevBayes

RevBayes (Höhna et al., 2016) is an open-source software program for Bayesian phylogenetic inference. It offers a flexible framework for hierarchical Bayesian modeling of complex and biologically realistic models of evolution. This flexibility is possible because RevBayes uses probabilistic graphical models (Höhna et al., 2014) and an interpreted programming language—called Rev—to specify and represent statistical models. For an review of the concepts and techniques used in this chapter, see Chapters 1.4 and 5.1 (Lartillot 2020; Pett and Heath 2020).

Links to RevBayes software and documentation

- Website: http://revbayes.com
- Download: http://revbayes.com/download
- Open source projects on GitHub: https://github.com/revbayes
- Tutorials: http://revbayes.com/tutorials
- Rev language reference: http://revbayes.com/documentation

In the probabilistic graphical modeling framework of RevBayes, model components (parameters and distributions) are interchangeable building blocks for constructing a complete statistical model (Höhna et al., 2016). This modularity enables users to easily modify a model to match their prior assumptions. When applying Bayesian analysis approaches, RevBayes uses a Markov chain Monte-Carlo (MCMC) algorithm to sample the posterior distributions of unknown parameters in a model. While inference using MCMC is the primary analysis approach in RevBayes, there are several other available statistical approaches, including model comparison using Bayes factors, and posterior predictive model checking and analysis of model adequacy (Höhna et al., 2018).

The core RevBayes library (written in C++) implements the various objects and functions that define a model and perform statistical analyses. Currently, the main interface to the RevBayes core is Rev, the interpreted programming language that users access via a RevBayes console or through writing Rev scripts. Members of the RevBayes Development Team are currently working to expand the set of interfaces for working with RevBayes and the Rev language. These include RevScripter[1] a graphical user interface for generating Rev analysis scripts, a Jupyter kernel[2] for running RevBayes in the Jupyter notebook environment, the RevKnitr[3] R package for using Rev interactively in RStudio, and the RevGadgets[4] R package for summarizing output from RevBayes analyses. Additional information on installing alternative graphical interfaces can be found on the RevBayes website[5].

The modular framework of RevBayes has facilitated the rapid expansion of available statistical methods for investigating evolutionary hypotheses. The tutorial presented here provides a mere glimpse at what is possible in RevBayes, focusing explicitly on inference of a time-calibrated phylogeny using paleontological and neontological data. However, there are a wide range of approaches for inferring macroevolutionary parameters in a phylogenetic framework. Throughout the tutorial, we refer to alternative or more advanced models and methods available in RevBayes. Thus, we hope that the exercises described here will introduce the reader to the potential for conducting analyses in RevBayes that may elucidate the evolutionary processes underpinning the generation of their biological data.

2 Background: Inferring the Timing and Phylogeny of Fossil and Extant Taxa

> This tutorial and associated files (i.e., data and script files) are maintained on the RevBayes website: `http://revbayes.com/tutorials/fbd_simple`.

The exercise described in Section 3 is a guide to using RevBayes to perform a simple phylogenetic analysis of extant and fossil bear species (family Ursidae), using morphological data as well as the occurrence times of lineages observed in the fossil record. To get an overview of the model, it is useful to think of the model as a generating process for our data. Suppose we would like to simulate our fossil and morphological data; we would consider two components (Figure 1):

- **Time tree model**: This is the diversification process that describes how a phylogeny is generated as well as when fossils are sampled along each lineage on the phylogeny. This component generates the phylogeny, divergence times, and the fossil occurrence data. The tree topology and node ages are parameters of the model that generates our morphological characters.
- **Discrete morphological character change model**: This model describes how discrete morphological character states change over time on the phylogeny. The generation of observed morphological character states is governed by other model components including

[1] RevScripter: `http://revbayes.com/revscripter`
[2] RevBayes Jupyter kernel: `https://github.com/revbayes/revbayes_kernel`
[3] RevKnitr: `https://github.com/revbayes/RevKnitr`
[4] RevGadgets R package `https://github.com/revbayes/RevGadgets`
[5] Graphical User Interface installation: `http://revbayes.com/gui-setup`

the substitution process and variation among characters in our matrix and among branches on the tree.

These two components, or modules, form the backbone of the inference model and reflect our prior beliefs on how the tree, fossil data, and morphological trait data are generated. We will provide a brief overview of the specific models used within each component while pointing to other tutorials that implement alternative models.

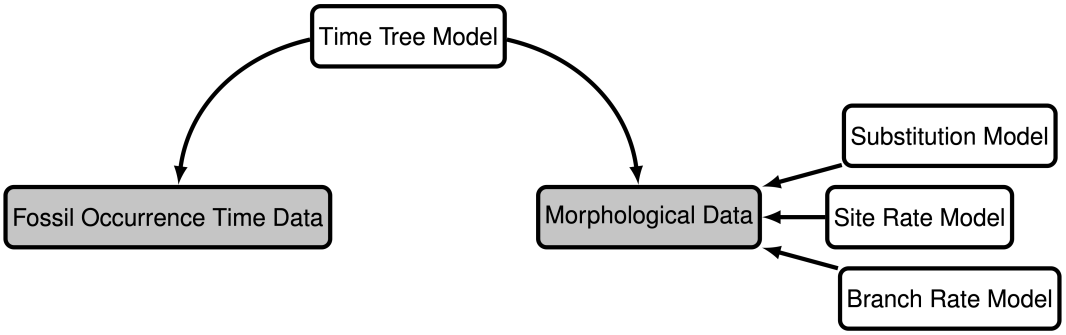


Figure 1 Modular components of the graphical model used in the analysis described in this tutorial. The gray boxes indicate the observed data: fossil ages and discrete morphological characters. The white boxes represent the models that generated the data. See also Section 5 of Chapter 5.1 [Pett and Heath 2020], and, in particular Figure 5, for other hierarchical models.

2.1 Time tree model: the fossilized birth-death process

The fossilized birth-death (FBD) process provides a joint distribution on the divergence times of living and extinct species, the tree topology, and the sampling of fossils (Stadler, 2010; Heath et al., 2014). The FBD model can be broken into two sub-processes, the birth-death process and the fossilization process.

2.1.1 Birth-death process

The birth-death process is a branching process that provides a distribution for the tree topology and divergence times on the tree. We will consider a constant-rate birth-death process (Kendall, 1948; Thompson, 1975). Specifically, we will assume every lineage has the same constant rate of speciation λ and rate of extinction μ at any moment in time (Nee et al., 1994; Höhna, 2015). Speciation and extinction events occur with rate parameters λ and μ respectively, whereby the waiting time between events is exponentially distributed with parameter $(\lambda + \mu)$. Then, given an event occurred, the probability of the event being a speciation is $(\lambda \ / \ (\lambda + \mu))$ while the probability of the event being an extinction is $(\mu \ / \ (\lambda + \mu))$.

The birth-death process depends on two other parameters as well, the origin time and the sampling probability. The origin time, denoted ϕ, represents the starting time of the stem lineage, which is the age of the entire process. The sampling probability, denoted ρ, gives the probability that an extant species is sampled.

The assumption that, at any given time, each lineage has the same speciation rate and extinction rate may not be realistic or valid in some systems. Several models are currently implemented in RevBayes that relax the assumption of constant rates such as,

episodic diversification rates[6] (Höhna, 2015), environment-dependent diversification rates[7] (Condamine et al., 2018), branch-specific diversification rates[8] (Höhna et al., 2019), or diversification rates tied to a species trait[9] (Maddison et al., 2007; Freyman and Höhna, 2018, 2019).

2.1.2 Fossilization process

Given a phylogeny, in this case a phylogeny generated by a birth-death process, the fossilization process provides a distribution for sampling fossilized occurrences of lineages in the tree (Heath et al., 2014). Much like speciation and extinction, fossil sampling is modeled according to a Poisson process with rate parameter ψ. This means that each lineage has the same constant rate of producing a fossil. As a result, along a given lineage, the time between fossilization events is exponentially distributed with rate ψ.

One key assumption of the FBD model is that each fossil represents a distinct fossil specimen. However, if certain taxa persist through time and fossilize particularly well, then the same taxon may be sampled at different stratigraphic ages. These fossil data are commonly represented by only the first and last appearances of a fossil morphospecies. In this case one might want to consider the fossilized birth-death range process[10] (Stadler et al., 2018) in RevBayes to model the stratigraphic ranges of fossil occurrences.

2.1.3 Accounting for fossil age uncertainty

Often, there is uncertainty around the age of each fossil, which is typically represented as an interval of the minimum and maximum possible ages. Moreover, a recent study demonstrated using simulated data that ignoring uncertainty in fossil occurrence dates can lead to biased estimates of divergence times (Barido-Sottani et al., 2019). RevBayes allows fossil occurrence time uncertainty to be modeled by directly treating it as part of the likelihood of the fossil data given the time tree. We model this by assuming the likelihood of a particular fossil occurrence $\mathcal{F}_i$ is zero if the inferred age t_i occurs outside the time interval (a_i, b_i) and some non-zero likelihood when the fossil is placed within the interval. Specifically, we will assume the fossil could occur anywhere within the observed interval with uniform probability, this means that the likelihood is equal to one if the inferred fossil age is consistent with the observed fossil interval:

$$f[\mathcal{F}_i \mid a_i, b_i, t_i] = \begin{cases} 1 & \text{if } a_i < t_i < b_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The incorporation of uncertainty around the fossil occurrence data is shown graphically as a part of our model in (Figure 2).

2.2 Modeling discrete morphological character change

Given a phylogeny, the discrete morphological character change model will describe how traits change along each lineage, resulting in the observed character states of fossils and

[6] Episodic diversification rates tutorial: `http://revbayes.com/tutorials/divrate/ebd`
[7] Environment-dependent diversification rates tutorial: `http://revbayes.com/tutorials/divrate/env`
[8] Branch-specific diversification tutorial: `http://revbayes.com/tutorials/divrate/branch_specific`
[9] State-dependent diversification tutorial: `http://revbayes.com/tutorials/sse/bisse-intro`
[10] Fossilized birth-death range process tutorial: `http://revbayes.com/tutorials/fbd_range`

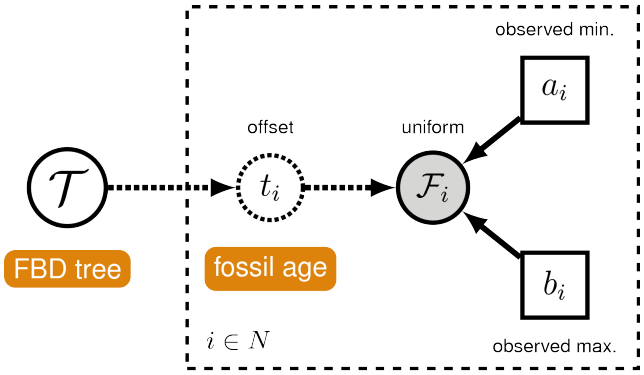


Figure 2 A graphical model of the fossil age likelihood model used in this tutorial. The likelihood of fossil observation $\mathcal{F}_i$ is uniform and non-zero when the inferred fossil age t_i falls within the observed time interval (a_i, b_i).

living species. In our case, the phylogeny and fossil occurrences are generated from the FBD process and we will be modeling the evolution of discrete morphological characters with two states. There are three main components to consider with modeling discrete morphological traits (as shown in Figure 1): the substitution model, the branch rate model, and the site rate model.

2.2.1 Substitution model

The substitution model describes how discrete morphological characters evolve over time. We will be using the Mk model (Lewis, 2001), a generalization of the Jukes-Cantor (Jukes and Cantor, 1969) model described for nucleotide substitutions (see Chapter 1.1 [Pupko and Mayrose 2020]).

The Mk model assumes that all transitions from one state to another occur at the same rate, for all k states. Since the characters used in this tutorial all have two states, we will specifically be using a model where $k = 2$. Thus, a transition from state 0 to state 1 is equally as likely as a transition from state 1 to state 0. For this tutorial, we focus on binary (2-state) characters for simplicity, but it is important to note that RevBayes can also accommodate multistate characters[11].

The evolution of discrete morphological characters is thought to occur at a very slow rate. Moreover, once some characters transition to a certain state, they rarely transition back, which means that the assumption of symmetric rates is likely violated my many empirical datasets (Wright et al., 2016; Wright, 2019). We can accommodate asymmetric transition rates[12] for each state using alternative models in RevBayes. Additionally, if some characters change symmetrically while others change asymmetrically, it is possible to partition[13] the matrix to account for model heterogeneity among characters.

[11] Multistate discrete morphology tutorial: `http://revbayes.com/tutorials/morph_tree/V2`
[12] Asymmetric transition rates tutorial: `http://revbayes.com/tutorials/morph_tree`
[13] Partitioned data analysis tutorial: `http://revbayes.com/tutorials/partition`

2.2.2 Branch-rate model

The branch-rate model describes how rates of morphological state transitions vary among branches in the tree. Each lineage in the phylogeny is assigned a value that acts as a scalar for the rate of character evolution. In our case we assume each branch has the same rate of evolution, this is a strict morphological clock (analogous to a strict molecular clock Zuckerkandl and Pauling, 1962). It is also possible to account for variation in rates among branches. These "relaxed-clock" models are commonly applied to molecular datasets and are currently implemented in RevBayes[14] (see Chapter 4.4 [Bromham 2020]).

2.2.3 Site-rate model

The rate of character evolution can often vary from site to site, i.e., from one column in the matrix to another (see Chapter 1.1 [Pupko and Mayrose 2020]). Under the site-rate model, a scalar is applied to each character to account for variation in relative rates. In our case we will assume that each character belongs to one of four rate categories from the discretized gamma distribution (Yang, 1994), which is parameterized by shape parameter α and number of rate categories n. Normally a gamma distribution requires shape α and rate β parameters, however, we set our site rates to have a mean of one, which results in the constraint $\alpha = \beta$, thus eliminating the second parameter. The parameter n breaks the gamma distribution into n equiprobable bins where the rate value of each bin is equal to its mean or median.

2.3 Putting together the complete phylogenetic model

We have outlined the specific components forming the processes that govern the generation of the time tree and morphological character data; and together these modules make up the complete phylogenetic model. Figure 3 shows the complete probabilistic graphical model that includes all of the parameters we will use in this tutorial (for more on graphical models for statistical phylogenetics see Höhna et al., 2014).

The parameters represented as stochastic nodes (solid white circles) in Figure 3 are unknown random variables that are estimated in our analysis. For each of these parameters, we assume a prior distribution that describes our uncertainty in that parameter's value. For example, we apply an exponential distribution with a rate of 10 as a prior on the mutation rate: $\mu \sim \text{Exponential}(10)$. The parameters represented as constant nodes (white boxes) are fixed to "known" or asserted values in the analysis.

2.4 Alternative models and analyses

The model choices and analysis in this tutorial focus on a simple example. Importantly, the modular design of RevBayes allows for many model choices to be swapped with more complex or biologically relevant processes for a given system. Analyses of a wide range of data types are also implemented in RevBayes (e.g., nucleotide sequences[15], historical biogeographic ranges[16]). Moreover, it is possible to fully integrate models describing the generation of data from different sources like in the "combined-evidence" approach[17] (Ronquist et al., 2012; Zhang et al., 2016; Gavryushkina et al., 2017) in a single, hierarchical Bayesian model. Some

[14] Relaxed clock models tutorial: `http://revbayes.com/tutorials/clocks`
[15] Nucleotide substitution models tutorial: `http://revbayes.com/tutorials/ctmc`
[16] Modeling discrete biogeography tutorial: `http://revbayes.com/tutorials/biogeo/biogeo_intro`
[17] FBD combined evidence tutorial: `http://revbayes.com/tutorials/fbd/fbd_specimen`

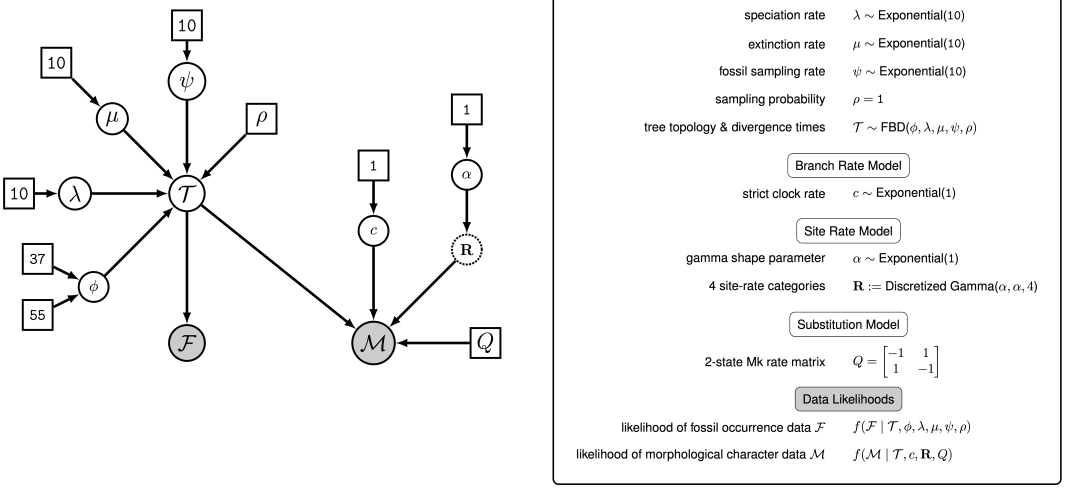


Figure 3 The complete graphical model used in the analysis described in this tutorial. This explicit representation of the model expands on the modular version depicted in Figure 1. The model components are defined in the box on the right. To simplify the model, we do not represent the components accounting for fossil age uncertainty illustrated in Figure 2.

researchers may wish to perform analyses with node calibrations[18], and this approach is also possible in RevBayes. Ultimately, for any statistical analysis of empirical data, it is important to consider the processes governing the generation of those data and how they can be represented in a hierarchical model.

3 Exercise: Phylogenetic Inference under the Fossilized Birth-Death Process

In this exercise, we will create a script in Rev, the interpreted programming language used by RevBayes, that defines the model outlined above and specifies the details of the MCMC simulation. This script can be executed in RevBayes to complete the full analysis. We conclude the exercise by evaluating the performance of the MCMC and summarizing the results.

3.1 Data and files

This tutorial is maintained online at `http://revbayes.com/tutorials/fbd_simple`. On this page, you will also find links to other RevBayes tutorials that will supplement some of the information provided here. Additionally, this site gives links to the data files and the completed script file.

[18] Molecular dating with node calibrations tutorial: `http://revbayes.com/tutorials/dating/nodedate`

On your own computer or your remote machine, create a directory called **RB_FBD_Tutorial** (or any name you like).

Then, navigate to the folder you just created and make a new directory called **data**.

In the **data** folder, add the following files (you can download these files from the URLs provided):

- **bears_taxa.tsv**: a tab-separated table listing the 18 bear species in our analysis (both fossil and extant) and their occurrence age ranges (minimum and maximum ages). For extant taxa, the minimum age is 0.0 (i.e., the present).
  http://revbayes.com/tutorials/fbd_simple/data/bears_taxa.tsv
- **bears_morphology.nex**: a matrix of 62 discrete, binary (coded **0** or **1**) morphological characters for our 18 species of fossil and extant bears.
  http://revbayes.com/tutorials/fbd_simple/data/bears_morphology.nex

Now you can create a separate file for the Rev script.

In the **RB_FBD_Tutorial** directory created above, create a blank file called **FBD_tutorial.Rev** and open it in a text editor.

It is also possible (though not recommended) to execute this entire tutorial in the RevBayes console.

The file **FBD_tutorial.Rev** will contain all of the instructions required to load the data, assemble the different model components used in the analysis, and configure and run the Markov chain Monte Carlo (MCMC) analysis. Once you finish writing this file, you can compare your script with the **FBD_tutorial.Rev** file on the tutorial webpage.

3.2 Importing data into RevBayes

We will begin our Rev script by loading in the two data files that were downloaded and saved to the **data** directory. In RevBayes, we use functions to read the contents of files and assign them to variables in our workspace. First, we will create a variable called **taxa** that will contain the data read in from **bears_taxa.tsv**.

```
taxa <- readTaxonData("data/bears_taxa.tsv")
```

Next, we will import the morphological character matrix from **bears_morphology.nex** and assign it to the variable **morpho**. In this exercise, we are using a NEXUS-formatted data file, but it is worth noting that several other file-types are acceptable depending on the kind of data (e.g., FASTA for molecular data).

```
morpho <- readDiscreteCharacterData("data/bears_morphology.nex")
```

Here, we use the function **readDiscreteCharacterData** to load a data matrix to the workspace from a formatted file. This function can be used for discrete morphological data as well as molecular sequence data (e.g., nucleotides, amino acids).

3.3 Helper variables

Before we begin specifying the hierarchical model, it is useful to instantiate some "helper variables" that will be used in our model and MCMC specification throughout our script.

First, we will create a new constant node called **n_taxa** that is equal to the number of species in our analysis (18).

```
n_taxa <- taxa.size()
```

Next, we will create a workspace variable called **moves**, which is a vector that will contain all of the MCMC moves used to propose new states for every stochastic node in the model graph. Each time a new stochastic node is created in the model, we can append the corresponding moves to this vector.

```
moves = VectorMoves()
```

One important distinction here is that **moves** is part of the RevBayes workspace and not the hierarchical model. Thus, we use the workspace assignment operator **=** instead of the constant node assignment operator **<-**.

3.4 The fossilized birth-death process

3.4.1 Speciation and extinction rates

Two key parameters of the FBD process are the speciation rate (the rate at which lineages are added to the tree, denoted by λ in Figure 3) and the extinction rate (the rate at which lineages are removed from the tree, μ in Figure 3). We will place exponential priors on both of these values, meaning we assume each parameter is drawn independently from a different exponential distribution, where each distribution has a rate parameter equal to **10**. Note that an exponential distribution with a rate of **10** has an expected value (mean) of **1/10**.

Create the exponentially distributed stochastic nodes for the **speciation_rate** and **extinction_rate** using the **~** stochastic assignment operator.

```
speciation_rate ~ dnExponential(10)
extinction_rate ~ dnExponential(10)
```

The **~** operator in Rev instantiates a stochastic node in the model (i.e., a solid circle in Figure 3). Every stochastic node must be defined by a distribution. In this case, we use the exponential. In the Rev language, every distribution has the prefix **dn** to make it easier to locate the various distributions in the Rev language documentation (http://revbayes.com/documentation). When a stochastic node is created in the model, the distribution function assigns it an initial value by drawing a random value from the prior distribution and assigns the node to the named variable.

For every stochastic node we declare, we must also specify proposal algorithms (called *moves*) to sample the value of the parameter in proportion to its posterior probability (see Chapter 1.4 [Lartillot 2020]). If a move is not specified for a stochastic node, then it will not be estimated, but fixed to its initial value.

The extinction rate and speciation rate are both positive, real numbers (i.e., non-negative floating point variables). For both of these nodes, we will use a scaling move (**mvScale**), which proposes multiplicative changes to a parameter.

```
moves.append(mvScale(speciation_rate, weight=1))
moves.append(mvScale(extinction_rate, weight=1))
```

You will also notice that each move has a specified **weight**. This option indicates the frequency a given move will be performed in each MCMC cycle. In RevBayes, the MCMC is executed by default with a *schedule* of moves at each step of the chain, instead of just one move per step, as is done in MrBayes (Ronquist and Huelsenbeck, 2003) or BEAST (Drummond et al., 2012; Bouckaert et al., 2014). Here, if we were to run our MCMC with our current vector of two moves each with a weight of **1**, then our move schedule would perform two moves in each cycle. Within a cycle, an individual move is chosen from the move list in proportion to its weight. Therefore, with both moves assigned **weight=1**, each has an equal probability of being executed and will be performed on average one time per MCMC cycle. For more information on moves and how they are performed in RevBayes, please refer to the tutorials introducing Markov chain Monte Carlo[19] and nucleotide substitution models[20].

In addition to the speciation (λ) and extinction (μ) rates, we may also be interested in inferring the net diversification rate ($\lambda - \mu$) and the turnover (μ/λ). Since these parameters can each be expressed as a deterministic transformation of the speciation and extinction rates, we can monitor their values (i.e., track their values and print them to a file) by creating two deterministic nodes using the **:=** deterministic assignment operator.

```
diversification := speciation_rate - extinction_rate
turnover := extinction_rate/speciation_rate
```

Deterministic nodes are represented by circles with dotted borders in a probabilistic graphical model. To maintain the simplicity of the model in Figure 3, the diversification rate and turnover are not shown.

3.4.2 Extant sampling probability

Every extant bear species is represented in this dataset. Therefore, we will fix the probability of sampling an extant lineage (ρ in Figure 3) to 1. The parameter **rho** will be specified as a constant node (new values for **rho** will not be sampled in the MCMC) using the **<-** constant assignment operator.

```
rho <- 1.0
```

Because ρ is a constant node, we do not have to assign a move to this parameter because we assume the value is known and fixed.

3.4.3 Fossil sampling rate

Since our data set includes serially sampled lineages, we must also account for the rate of sampling through time. This is the fossil sampling (or recovery) rate (ψ in Figure 3), which we will instantiate as a stochastic node named **psi**. As with the speciation and extinction rates (see Section 3.4.1), we will use an exponential prior on this parameter and apply a scale move to sample values from the posterior distribution.

```
psi ~ dnExponential(10)
moves.append(mvScale(psi, weight=1))
```

[19] Introduction to MCMC tutorial: http://revbayes.com/tutorials/mcmc/
[20] Nucleotide substitution models tutorial: http://revbayes.com/tutorials/ctmc

3.4.4 Origin time

The FBD process is conditioned on the origin time (ϕ in Figure 3), which requires specification of a node representing the age of the clade. We will set a uniform distribution on the origin age, with the lower bound set at the age of the oldest bear fossil (37 My) and the higher bound of 55 My set to the age of the most recent common ancestor of crown Carnivora estimated by recent studies (dos Reis et al., 2012). For the move, we will use a sliding window move (**mvSlide**), which samples a parameter uniformly within an interval (defined by the half-width "delta", which is set to **1** by default). Sliding window moves can be problematic for small values, as the window may overlap zero. However, our prior on the origin age excludes values ≤ 37.0, so this is not an issue.

```
origin_time ~ dnUnif(37.0, 55.0)
moves.append(mvSlide(origin_time, weight=1.0))
```

3.4.5 The FBD tree

Now that we have specified all of the parameters of the FBD process (λ, μ, ϕ, ψ), we will use these parameters to create the stochastic node representing the time-calibrated tree that we will call **fbd_tree**. The **fbd_tree** ($\mathcal{T}$ in Figure 3) is generated by a fossilized birth-death distribution and is conditionally dependent on λ, μ, ϕ, and ψ. The FBD distribution function **dnFBDP** takes the FBD parameters as arguments as well as the **taxa** variable which specifies the number of terminal taxa as well as the taxon labels.

```
fbd_tree ~ dnFBDP(origin=origin_time, lambda=speciation_rate,
                  mu=extinction_rate, psi=psi, rho=rho, taxa=taxa)
```

Next, in order to sample from the posterior distribution of trees, we need to specify moves that propose changes to the topology (**mvFNPR**) and node times (**mvNodeTimeSlideUniform**). We also include a proposal (**mvCollapseExpandFossilBranch**) that will collapse or expand a fossil branch, thus sampling trees where a given fossil is either a sampled ancestor or a sampled tip. In addition, when conditioning on the origin time, we also need to explicitly sample the root age (**mvRootTimeSlideUniform**).

```
moves.append(mvFNPR(fbd_tree, weight=15.0))
moves.append(mvCollapseExpandFossilBranch(fbd_tree, origin_time,
                                          weight=6.0))

moves.append(mvNodeTimeSlideUniform(fbd_tree, weight=40.0))
moves.append(mvRootTimeSlideUniform(fbd_tree, origin_time,
                                    weight=5.0))
```

Note that we specified a higher move **weight** for each of the proposals operating on **fbd_tree** than we did for the previous stochastic nodes. This means that our move schedule will propose fifteen times as many new topologies via the **mvFNPR** move as it will new values of **speciation_rate** using **mvScale**, for example. By increasing the number of times new values are proposed for a parameter, we are increasing the sampling intensity for that parameter. Typically, we do this for parameters that we are particularly interested in or for parameters that tend to induce long mixing times. A node like $\mathcal{T}$ in our graphical model (Figure 3) represents a complex set of variables: the tree topology and all divergence times. Moreover, the likelihoods of our fossil occurrence data and the morphological character data are both conditionally dependent on the time tree. Such complex variables require more extensive sampling than other nodes.

3.4.6 Sampling fossil cccurrence times

We need to account for uncertainty in the age estimates of our fossils using the observed minimum and maximum stratigraphic ages that are provided in the file **bears_taxa.tsv**. We can represent the fossil likelihood using any uniform distribution that is non-zero when the likelihood is equal to one (see Section 2.1.3). For example, if t_i is the inferred fossil age and (a_i, b_i) is the observed stratigraphic interval, we know the likelihood is equal to one when $a_i < t_i < b_i$, or equivalently $t_i - b_i < 0 < t_i - a_i$. So we can represent this likelihood using a uniform random variable, uniformly distributed in $(t_i - b_i, t_i - a_i)$ and clamped at zero.

To do this, we will get all the fossils from the tree and use a **for** loop to iterate over them. For each fossil observation, we will create a uniform random variable representing the likelihood, based on the minimum and maximum ages specified in the file **bears_taxa.tsv**.

```
fossils = fbd_tree.getFossils()
for(i in 1:fossils.size())
{
    t[i] := tmrca(fbd_tree, clade(fossils[i]))

    a_i = fossils[i].getMinAge()
    b_i = fossils[i].getMaxAge()

    F[i] ~ dnUniform(t[i] - b_i, t[i] - a_i)
    F[i].clamp( 0 )
}
```

Finally, we will add a move that samples the ages of all the fossils on the tree.

```
moves.append(mvFossilTimeSlideUniform(fbd_tree, origin_time,
                                      weight=5.0))
```

3.4.7 Monitoring parameters of interest

There are additional parameters that may be of particular interest to us that are not directly sampled as part of the graphical model defined thus far. As with the diversification and turnover nodes specified in Section 3.4.1, we can create deterministic nodes to sample the posterior distributions of these parameters. Here we will create a deterministic node called **num_samp_anc** that will compute the number of sampled ancestors in our **fbd_tree**.

```
num_samp_anc := fbd_tree.numSampledAncestors()
```

We are also interested in the age of the most-recent-common ancestor (MRCA) of all living bears. To monitor this age in our MCMC sample, we must use the **clade()** function to identify the node corresponding to the MRCA. Once this clade is defined we can instantiate a deterministic node called **age_extant** that will record the age of the MRCA of all living bears, using the **tmrca()** function.

```
clade_extant = clade("Ailuropoda_melanoleuca","Tremarctos_ornatus",
                     "Melursus_ursinus","Ursus_arctos",
                     "Ursus_maritimus","Helarctos_malayanus",
                     "Ursus_americanus","Ursus_thibetanus")
age_extant := tmrca(fbd_tree, clade_extant)
```

In the same way we monitored the MRCA of the extant bears, we can also monitor the age of a fossil taxon that we may be interested in recording. We will monitor the marginal

distribution of the age of *Kretzoiarctos beatrix* (Abella et al., 2012), which is sampled between 11.2–11.8 My.

```
age_Kretzoiarctos_beatrix := tmrca(fbd_tree,
                                   clade("Kretzoiarctos_beatrix"))
```

3.5 Modeling the evolution of binary morphological characters

The next part of the graphical model, we will define specifies the model of morphological character evolution. This component includes the substitution model, the model of rate variation among characters, and the model of rate variation among branches (Figure 3).

As stated in Section 2.2.1, we will use the Mk model to describe the substitution process. Because the Mk model is a generalization of the Jukes-Cantor model (Jukes and Cantor, 1969), we will initialize our instantaneous rate matrix from a Jukes-Cantor matrix (see Chapter 1.1 [Pupko and Mayrose 2020]). The constant node **`Q_morpho`** corresponds to the two-state rate matrix Q in Figure 3.

```
Q_morpho := fnJC(2)
```

We will assume that rates vary among characters in our data matrix according to a discretized gamma distribution (described in Section 2.2.3). For this model, we create a vector of rates named **`rates_morpho`** which is the product of a function **`fnDiscretizeGamma()`** that divides up a gamma distribution into a set of equal-probability bins (**R** in Figure 3). Here, our only stochastic node is **`alpha_morpho`** (α in Figure 3), which is the shape parameter of the discretized gamma distribution.

```
alpha_morpho ~ dnExponential(1.0)
rates_morpho := fnDiscretizeGamma(alpha_morpho, alpha_morpho, 4)

moves.append(mvScale(alpha_morpho, weight=5.0))
```

The phylogenetic model also assumes that each branch has a rate of morphological character change. For simplicity, we will assume a strict morphological clock–meaning that every branch has the same rate represented by the stochastic node **`clock_morpho`** (c in Figure 3), which is drawn from an exponential distribution (see Section 2.2.2).

```
clock_morpho ~ dnExponential(1.0)
moves.append(mvScale(clock_morpho, weight=4.0))
```

3.5.1 The phylogenetic CTMC

If you refer to Figure 3, you will see that we have defined almost all of the components of the complete model except for the observed node representing our morphological character data ($\mathcal{M}$). The character matrix is a clamped stochastic node that is generated by a phylogenetic continuous-time Markov chain (CTMC) distribution (see Chapter 1.1 [Pupko and Mayrose 2020]). This node is conditionally dependent on the time tree ($\mathcal{T}$: **`fbd_tree`**), clock rate (c: **`clock_morpho`**), site rates (**R**: **`rates_morpho`**), and the two-state Mk rate matrix (Q: **`Q_morpho`**). With all of these nodes instantiated in the graphical model, we can now connect the components by defining the node representing our observed morphological data.

There are some unique aspects to specifying a phylogenetic CTMC for morphological data. You will notice that we have an option called **`coding`**. This option allows us to condition on

biases in the way the morphological data were collected (i.e., ascertainment bias). By setting **coding=variable** we can correct for coding only variable characters (as discussed in Lewis, 2001).

```
phyMorpho ~ dnPhyloCTMC(tree=fbd_tree, siteRates=rates_morpho,
                        branchRates=clock_morpho, Q=Q_morpho,
                        type="Standard", coding="variable")
phyMorpho.clamp(morpho)
```

Now that we have defined our complete model, we can create a workspace variable that packages the entire model graph. This makes it easy to pass the whole model to functions that will set up our MCMC analysis. This variable is created using the **model()** function, which takes only a single node in the graph. We will use the **fbd_tree** node, but you can try this with an alternative node (e.g., **clock_morpho**, **rho**, etc.). As long as you have established all of the connections among the model parameters, the **model()** function will find every other node by traversing the edges of the graph (Figure 3).

```
mymodel = model(fbd_tree)
```

3.6 Monitoring variables

We have defined the full probabilistic graphical model shown in Figure 3 and now we are ready to specify the details of our MCMC analysis. The first step in setting up the analysis is to create *monitors* that will record the values of each parameter in our model for every sampled cycle of the MCMC. The sampled values are saved to file (or printed to screen) and can be summarized when our MCMC simulation is complete.

Let's create three different monitor objects for this analysis. To manage the monitors in RevBayes, we create another workspace variable called **monitors** that is a vector containing the three monitor variables.

```
monitors = VectorMonitors()
```

We will append our first monitor to the **monitors** vector. This will create a file called **bears.log** in a directory called **output** (if this directory does not already exist, RevBayes will create it). The function **mnModel()** initializes a monitor that saves all of the numerical parameters in the model to a tab-delineated file. This file is useful for summarizing marginal posteriors in statistical plotting tools like Tracer (Rambaut et al., 2018) or R (R Core Team, 2020). We will exclude the **F** vector from logging, as it is purely used as an auxiliary variable for estimating fossil ages, and is clamped to 0. Additionally, we also specify how frequently we sample our Markov chain by setting the **printgen** option. We will sample every **10** cycles of our MCMC.

```
monitors.append(mnModel(filename="output/bears.log", printgen=10,
                        exclude=["F"]))
```

You may think that sampling every 10 generations may be too frequent to avoid correlation between samples in our MCMC. However, recall that a single "generation" in RevBayes performs a schedule of moves that is determined by the number of moves in the **moves** vector and the weights assigned to those moves (see Section 3.4.1). Thus, a single generation in this analysis will involve 26 moves, so if we record every 10 generations, there will be 260 moves between each sample.

We want to create a separate file containing samples of the tree and branch lengths since these will not be saved by the monitor defined above. To save the tree parameter, we can use the **mnFile()** function that saves specific parameters to a file. We indicate the parameters by including them in the function's options.

```
monitors.append(mnFile(filename="output/bears.trees", printgen=10,
                       fbd_tree))
```

The final monitor will print updates of our MCMC to the screen. The screen monitor function **mnScreen()** allows us to add parameters in our model that will be displayed along with a few default values (including the current iteration, posterior, likelihood, and prior). We will monitor the age of the MRCA of the living bears, the number of sampled ancestors, and the origin time in the screen output.

```
monitors.append(mnScreen(printgen=10, age_extant, num_samp_anc,
                         origin_time))
```

3.7 Setting up and running the MCMC sampler

Our Rev script specifies the three major parts of our MCMC analysis: a model (**mymodel**), a list of MCMC proposals (**moves**), and a way to save the values sampled by our Markov chain (**monitors**). With these three components, we can set up our analysis using the **mcmc()** function. This function creates a workspace variable that we can use to execute the MCMC simulation.

```
mymcmc = mcmc(mymodel, monitors, moves)
```

Using our variable **mymcmc**, we can execute the **run()** member method to start our MCMC sampler.

```
mymcmc.run(generations=10000)
```

Finally, since we are going to save this analysis in a script file and run it in RevBayes, it is useful to include a statement that will quit the program when the run is complete.

```
q()
```

Your script is now complete! Note that you can compare your script to the **FBD_tutorial.Rev** file provided on the tutorial webpage.

Save the **FBD_tutorial.Rev** file in the **RB_FBD_Tutorial** directory.

3.8 Execute the analysis script in RevBayes

With your script complete and data files in the proper location, you can execute the **FBD_tutorial.Rev** script in RevBayes.

Run the RevByes executable.

On Unix systems, if the RevBayes is in your path, you simply need to navigate to the **RB_FBD_Tutorial** directory and type **rb**.

If the RevBayes executable is not in your path, you can execute it and then change your working directory within the program using the **setwd()** function which takes the absolute path to your directory as an argument.

```
setwd("<path to>/RB_FBD_Tutorial")
```

Once RevBayes is in the correct working directory (**RB_FBD_Tutorial**), you can then use the **source()** function to feed RevBayes your master script file (**FBD_tutorial.Rev**).

```
source("FBD_tutorial.Rev")
```

This will execute the analysis and you should see the various parameters—specified when you initialized the screen monitor—printed to the screen every 10 generations. When the analysis is complete, RevBayes will quit and you will have a new directory called output that will contain all of the files you specified with the monitors.

3.9 Results

Two files are created by the monitors in Section 3.6. These files, located in the **output** directory contain the record of values sampled for the various parameters of the model over the course of the MCMC. In the following sections, we will assess the performance of our MCMC sampler and summarize the marginal posterior distributions of numerical parameters (in the file **bears.log**) and the time-calibrated phylogeny (in the file **bears.trees**).

3.9.1 Evaluating the MCMC sampler

The first step when analyzing the output of an MCMC run is to check whether the chain has converged on the stationary distribution and sampled effectively (i.e., achieved "good mixing"). This can be done by loading the parameter log, in our case the file **bears.log**, in a program such as Tracer[21] (Rambaut et al., 2018), shown in Figure 4.

On the left side is a panel summarizing all the parameters appearing in the log, with their mean estimate and ESS value (effective sample size). The ESS of a parameter determines whether the chain has adequately sampled the associated variable: values above 200 are considered "good", whereas values below 200, highlighted by Tracer in yellow or red, indicate poor mixing. Explicitly, the ESS measures the degree of independence between samples and parameters with signatures of autocorrelation between samples are indicative of an inadequate sampler.

Here we can see that the chain has mixed well for some parameters, but not others. In particular, we see low ESS values for the origin time (**origin_time**) and the ages of some fossil tips (**t[1]**, **t[9]** and **t[10]**). This may indicate that the MCMC sampler has not converged on the stationary distribution for these parameters, which are associated with the FBD tree. What this assessment reveals is that we did not perform enough proposals for these parameters. Thus, it will be important to run the MCMC for more generations

[21] Tracer: http://beast.community/tracer

5.2:18 Estimating a time-calibrated phylogeny of fossil and extant taxa using RevBayes

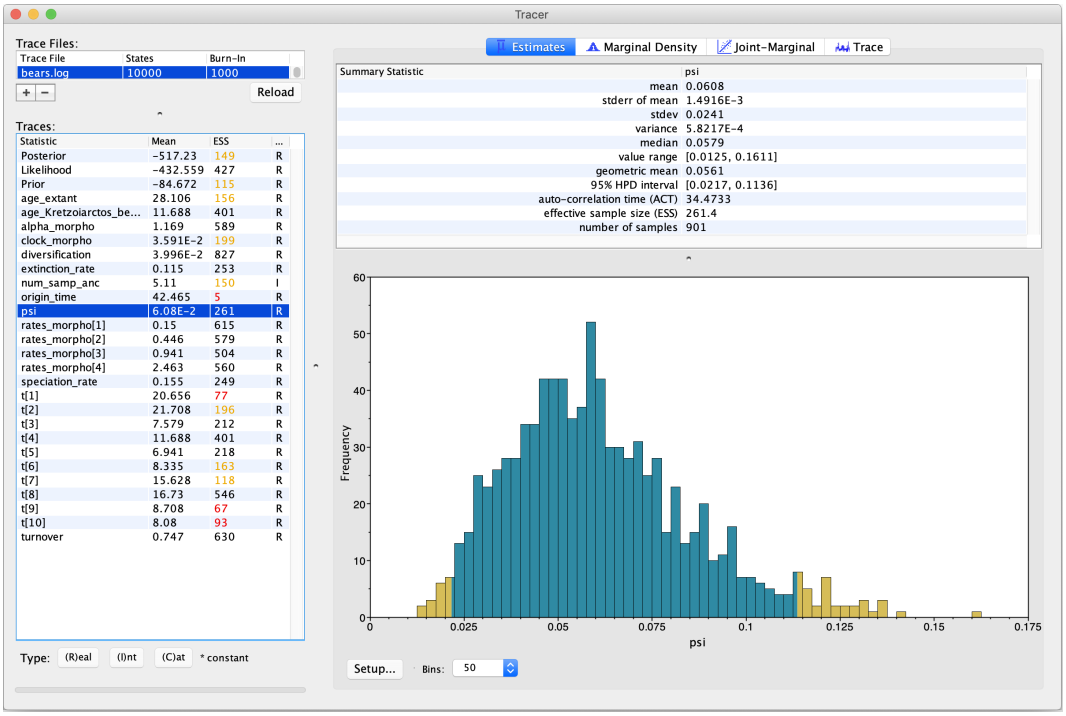


Figure 4 Analysis in Tracer of the parameter estimates obtained on the bears dataset.

(specified in Section 3.7) and/or increase the weights of moves applied to these stochastic nodes (e.g., the **mvSlide** applied to **origin_time** in Section 3.4.4). For more details on diagnosing convergence of MCMC samples under the FBD model, please see the tutorial on combined-evidence analysis in RevBayes [22].

3.9.2 Summarizing the tree

Once we are certain that our MCMC has effectively sampled the joint posterior distribution of our model parameters, we can summarize the tree topology, branch times, and fossil ages that were saved to **output/bears.trees** using some built-in RevBayes functions.

> Run the RevByes executable, making sure that the working directory is **RB_FBD_Tutorial**.

The file **bears.trees** contains the trees and associated parameters that were sampled every 10 generations by our monitor. In RevBayes, we often refer to a set of samples from our MCMC as a "trace".

Begin by loading the tree trace into RevBayes from the **bears.trees** file.

```
trace = readTreeTrace("output/bears.trees")
```

By default, a burn-in of 25% is used when reading in the tree trace (250 trees in our case). Note that this is different from Tracer, which uses a burn-in fraction of 10% by

[22] FBD combined evidence tutorial: http://revbayes.com/tutorials/fbd/fbd_specimen

default. You can specify a different burn-in fraction, say 50%, by typing the command `trace.setBurnin(500)`.

Now we will use the `mccTree()` function to return a maximum clade credibility (MCC) tree. The MCC tree is the tree with the maximum product of the posterior clade probabilities. When considering trees with sampled ancestors, we refer to the maximum sampled ancestor clade credibility (MSACC) tree (Gavryushkina et al., 2017).

```
mccTree(trace, file="output/bears.mcc.tre")
```

When there are sampled ancestors present, visualizing the tree can be fairly difficult in traditional tree viewers. We will make use of a browser-based tree visualization tool called IcyTree (Vaughan, 2017), which can be accessed at `https://icytree.org`. IcyTree has many unique options for visualizing phylogenetic trees and can produce publication-quality vector image files (i.e., SVG). Additionally, it correctly represents sampled ancestors on the tree as nodes, each with only one descendant (Figure 5).

Navigate to `https://icytree.org` and open the file **`output/bears.mcc.tre`** in IcyTree.

Try to replicate the tree in Figure 5 (Hint: ***Style > Mark Singletons***).

★ Why might a node with a sampled ancestor be referred to as a singleton?

★ How can you see the names of the fossils that are putative sampled ancestors?

★ What is the posterior probability that *Zaragocyon daamsi* is a sampled ancestor?

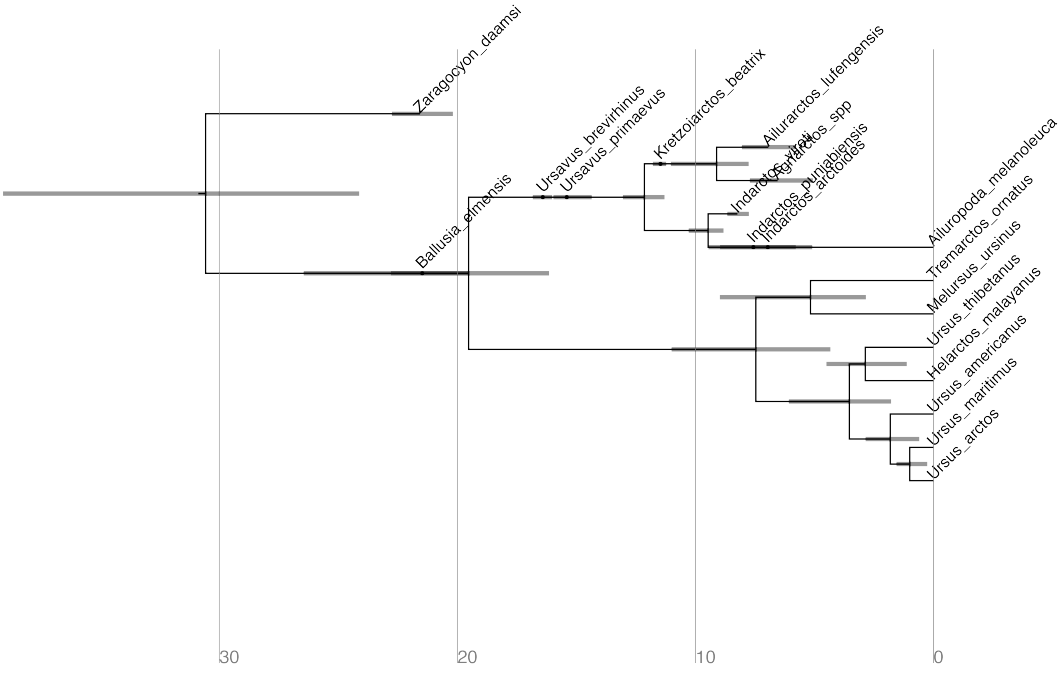


Figure 5 Maximum sampled ancestor clade credibility (MSACC) tree of bear species used in this tutorial.

3.10 Summary

In this tutorial, we have introduced core information about how morphological and age information are modeled for use with the FBD model in RevBayes. We have also discussed important aspects of executing and summarizing MCMC analysis. This exercise uses a simplified data set and set of models for analysis of fossil and extant data. Most researchers working on living taxa have access to molecular (including genomic) data and may be interested in applying these methods to much larger datasets and more complex problems. Note that the goal of this tutorial is to provide a concise introduction to the framework for analysis of paleontological and neontological data in RevBayes. For more information on how to apply RevBayes datasets combining morphological and molecular characters, please refer to the tutorial describing this approach: http://revbayes.com/tutorials/fbd/fbd_specimen.

4 Bayesian Phylogenetic Inference in RevBayes

This tutorial provided a very focused look at the range of models and methods available in RevBayes. There are currently numerous approaches available and under active development by RevBayes team members. These include (but are not limited to):

- Model selection using Bayes factors
- Model averaging of substitution models
- Approaches for assessing model adequacy using posterior prediction
- Analysis of multi-state discrete morphological characters under asymmetric models
- Various relaxed-clock models
- Models that vary diversification over time
- State-dependent diversification models
- Analysis of chromosome evolution
- Lineage specific diversification rate variation
- Analysis of continuous characters under Brownian motion and Ornstein-Uhlenbeck models
- Ancestral area estimation and phylogenetic analysis of historical biogeography
- Gene-tree/species-tree inference under the multi-species coalescent

The flexibility of the modeling framework implemented in RevBayes provides a rich tool-kit for phylogenetic analysis under complex models. Moreover, the RevBayes core and probabilistic graphical models make it possible for new developers to readily implement their ideas in an existing code base. Members of the RevBayes Development Team are working to expand the documentation for new developers (http://revbayes.com/developer) to facilitate the growth of new statistical models and methods available in RevBayes.

Acknowledgements

We wish to thank the editors C. Scornavacca, F. Delsuc, and N. Galtier, for the opportunity to contribute this tutorial to *Phylogenetics in the Genomic Era*. We also thank S. Höhna for providing comments on this manuscript. All RevBayes tutorials benefit from the generous feedback provided by researchers applying RevBayes and workshop participants. The methods described in this tutorial are available because of the efforts of the RevBayes Developer Team, a collaborative network of scientific programmers working on phylogenetic problems.

References

Abella, J., Alba, D. M., Robles, J. M., Valenciano, A., Rotgers, C., Carmona, R., Montoya, P., and Morales, J. (2012). *Kretzoiarctos* gen. nov., the oldest member of the giant panda clade. *PLoS One*, 17:e48985.

Barido-Sottani, J., Aguirre-Fernández, G., Hopkins, M. J., Stadler, T., and Warnock, R. (2019). Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth–death process. *Proceedings of the Royal Society B: Biological Sciences*, 286(1902):20190685.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4):e1003537.

Bromham, L. (2020). Substitution rate analysis and molecular evolution. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.4, pages 4.4:1–4.4:21. No commercial publisher | Authors open access book.

Condamine, F. L., Rolland, J., Höhna, S., Sperling, F. A., and Sanmartín, I. (2018). Testing the role of the Red Queen and Court Jester as drivers of the macroevolution of Apollo butterflies. *Systematic Biology*, 67(6):940–964.

dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C., and Yang, Z. (2012). Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences*, 279(1742):3491–3500.

Drummond, A., Suchard, M., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29:1969–1973.

Freyman, W. A. and Höhna, S. (2018). Cladogenetic and anagenetic models of chromosome number evolution: a Bayesian model averaging approach. *Systematic Biology*, 67(2):1995–215.

Freyman, W. A. and Höhna, S. (2019). Stochastic character mapping of state-dependent diversification reveals the tempo of evolutionary decline in self-compatible Onagraceae lineages. *Systematic Biology*, 68(3):505519.

Gavryushkina, A., Heath, T. A., Ksepka, D. T., Stadler, T., Welch, D., and Drummond, A. J. (2017). Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic Biology*, 66:57–73.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29):E2957–E2966.

Höhna, S. (2015). The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events. *Journal of Theoretical Biology*, 380:321–331.

Höhna, S., Coghill, L. M., Mount, G. G., Thomson, R. C., and Brown, J. M. (2018). P^3: Phylogenetic posterior prediction in RevBayes. *Molecular Biology and Evolution*, 35(4):1028–1034.

Höhna, S., Freyman, W. A., Nolen, Z., Huelsenbeck, J. P., May, M. R., and Moore, B. R. (2019). A Bayesian approach for estimating branch-specific speciation and extinction rates. *bioRxiv*, https://doi.org/10.1101/555805.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., and Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic Biology*, 63(5):753–771.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, 3:21–132.

Kendall, D. G. (1948). On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, 19(1):1–15.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6):913–925.

Maddison, W., Midford, P., and Otto, S. (2007). Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56(5):701.

Nee, S., May, R. M., and Harvey, P. H. (1994). The Reconstructed Evolutionary Process. *Philosophical Transactions: Biological Sciences*, 344(1309):305–311.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, 67(5):901–904.

Ronquist, F. and Huelsenbeck, J. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., and Rasnitsyn, A. P. (2012). A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology*, 61(6):973–999.

Stadler, T. (2010). Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*, 267(3):396–404.

Stadler, T., Gavryushkina, A., Warnock, R. C., Drummond, A. J., and Heath, T. A. (2018). The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. *Journal of Theoretical Biology*, 447:41–55.

Thompson, E. A. (1975). *Human Evolutionary Trees*. Cambridge University Press, Cambridge, UK.

Vaughan, T. G. (2017). IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, 33(15):2392–2394.

Wright, A. M. (2019). A systematist's guide to estimating Bayesian phylogenies from morphological data. *Insect systematics and diversity*, 3(3):2.

Wright, A. M., Lloyd, G. T., and Hillis, D. M. (2016). Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, 65(4):602–611.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. (2016). Total-evidence dating under the fossilized birth-death process. *Systematic Biology*, 65(2):228–249.

Zuckerkandl, E. and Pauling, L. (1962). Molecular disease, evolution, and genetic heterogeneity. In Kasha, M. and Pullman, B., editors, *Horizons in Biochemistry*, pages 189–225. Academic Press, New York.

Chapter 5.3 Efficiently Analysing Large Viral Data Sets in Computational Phylogenomics

Anna Zhukova
Unité Bioinformatique Evolutive, Hub Bioinformatique et Biostatistique, USR3756 (C3BI/DBC), Institut Pasteur & CNRS, Paris, France
anna.zhukova@pasteur.fr

Olivier Gascuel
Unité Bioinformatique Evolutive, USR3756 (C3BI/DBC), Institut Pasteur & CNRS, Paris, France

Sebastián Duchêne
Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, University of Melbourne, Australia

Daniel L. Ayres
Center for Bioinformatics and Computational Biology, University of Maryland, USA

Philippe Lemey[1]
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven – University of Leuven, Leuven, Belgium

Guy Baele[2]
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven – University of Leuven, Leuven, Belgium
guy.baele@kuleuven.be

Abstract

Viral evolutionary analyses are confronted with increasingly large sequence data sets, both in terms of sequence length and number of sequences. This can result in considerable computational burden, not only to infer phylogenies but also to obtain associated estimates such as their time scales and phylogeographic patterns. Here, we illustrate two frequently-used approaches to obtain phylogenomic estimates of time-measured trees and spatial dispersal patterns for fast-evolving viruses. First, we discuss computationally efficient procedures that employ a fixed tree topology obtained through maximum likelihood inference to estimate molecular clock rates and phylogeographic spread for Dengue virus genomes. Using the same viral example, we also illustrate Bayesian phylodynamic inference that jointly infers time-measured trees and phylogeography, including covariates of spatial dispersal, from sequence and trait data. We highlight state-of-the-art efforts to perform such computations more efficiently. Finally, we compare the estimates obtained by both approaches and discuss their strengths and potential pitfalls.

How to cite: Anna Zhukova, Olivier Gascuel, Sebastián Duchêne, Daniel L. Ayres, Philippe Lemey, and Guy Baele (2020). Efficiently Analysing Large Viral Data Sets in Computational Phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the*

[1] PL acknowledges support by the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G066215N, G0D5117N and G0B9317N).
[2] GB acknowledges support from the Interne Fondsen KU Leuven / Internal Funds KU Leuven under grant agreement C14/18/094, and the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G0E1420N).

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 5.3; pp. 5.3:1–5.3:43

A book completely handled by researchers.
No publisher has been paid.

Genomic Era, chapter No. 5.3, pp. 5.3:1–5.3:43. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Acknowledgements This work was supported by the EU-H2020 Virogenesis project (grant number 634650), by the INCEPTION project (PIA/ANR-16-CONV-0005), and by the Reservoir-DOCS ERC grant (agreement no. 725422). The Artic Network receives funding from the Wellcome Trust through project 206298/Z/17/Z.

1 Introduction

According to a "quick guide" to phylogenomics (Telford, 2007), standard phylogenomic approaches leverage the information present in a large number of genes generated in large-scale genome sequencing efforts. In infectious disease research, phylogenomic alignments that comprise single-nucleotide polymorphism (SNP) data of several hundreds or thousands of genes are now also an important focus of modern microbiology studies. However, for rapidly evolving RNA viruses, phylogenetic and evolutionary analyses remain inherently restricted to small genomes that encode for a limited number of genes. In this chapter, we focus on current challenges and opportunities for evolutionary inference from rapidly evolving viral genomes. This goes beyond common phylogenomic approaches and it is perhaps more in line with some of the earliest published mentions of phylogenomics that refer to a mixed bag of gene or genome analyses within a phylogenetic framework. We will primarily focus on the many different types of analyses that are being used in epidemiology, which shares many common interests with the field of phylodynamics.

The mention of "large" viral data sets in our title refers to the increase in two dimensions of the information available for studying molecular epidemiology and virus evolution, which has been brought about by the revolution in sequencing technology. The first dimension concerns the transition from a single gene or a typical PCR amplicon sequenced by Sanger sequencing to complete genomes that are now easily obtained through next generation sequencing, which roughly represents an increase of one order of magnitude for many RNA viruses. This seems less impressive than the increase from a single gene to hundreds of genes that phylogenomic studies of many other organisms have to confront. In addition, due to evolutionary rates that are about a million times faster than our own cellular genes, a limited marker in an RNA virus genome may already offer reasonable resolution for reconstructing evolutionary histories with time-scales of a decade or older. For example, a polymerase gene fragment of about 1 000 bp that is routinely sequenced for drug resistance testing has been extensively and successively used in HIV molecular epidemiology (e.g. Hué et al. 2005) while a fragment of less than half this size – but for the more variable envelope gene – has been used to reconstruct the origin and spread of the virus in Central Africa (Faria et al., 2014). Nevertheless, complete genomes offer an important increase in the resolution of the inferred phylogenies (Yebra et al., 2016), which for short-term outbreak dynamics in particular opens up new opportunities for epidemic reconstructions and tracking transmission. We illustrate this increase of phylogenetic resolution by comparing maximum likelihood trees for complete genome data and the corresponding glycoprotein gene sequences for the 2013 – 2016 West African Ebola virus outbreak in Figure 1. In this case, a single gene does not provide sufficient information about clustering of Ebola virus isolates whereas complete genomes do offer reasonable phylogenetic resolution allowing to identify some degree of structuring by country of sampling.

Complete genome sequencing has in recent years become the standard in outbreak

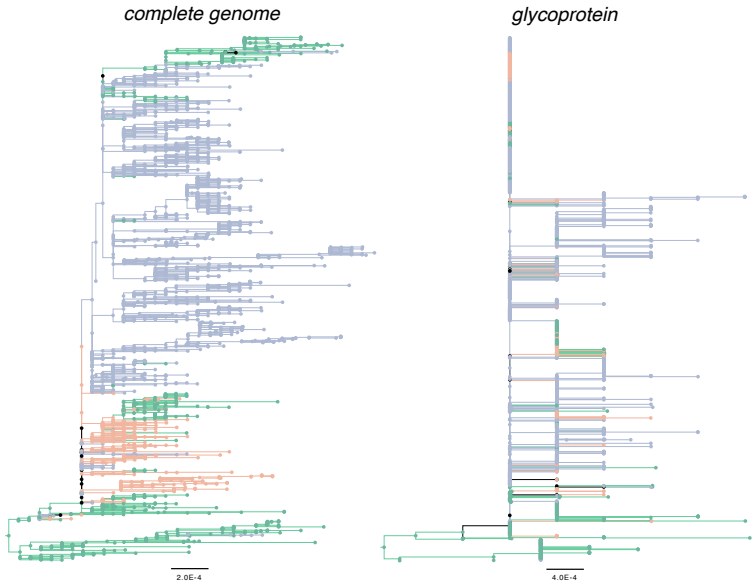


Figure 1 Maximum likelihood phylogenetic trees of Ebola virus inferred using complete genomes (left) and only the glycoprotein gene (right). Branches are coloured according to country (green: Guinea; blue: Sierra Leone; red: Liberia), based on a parsimony reconstruction for internal nodes. Complete genome data allow to infer reasonably resolved phylogenetic trees with a discernible structuring of lineages by country, whereas a single gene produces a multitude of polytomies and essentially prevents from uncovering any relevant (geographic) structure in the resulting phylogeny.

surveillance as illustrated in recent work on Zika (Faria et al., 2017), Chikungunya (Naveca et al., 2019) and Yellow Fever (Faria et al., 2018). Such complete genome data also come with specific challenges, such as the need to take into account recombination in different viruses or the difficulty to combine segments for viruses that undergo reassortment, or with the general challenge of increased computation times in likelihood-based phylogenetics (Chapter 1.2 [Stamatakis and Kozlov 2020]). In this chapter, we devote a great deal of attention to describing the methods that can be employed to address the computational challenges caused by these increases in data set sizes and model complexity. Related to this, Chapter 5.4 (Ayres et al. 2020) describes improvements in the latest version of the BEAGLE library for high-performance likelihood computation (Ayres et al., 2019), which enables parallel calculation of independent data partitions on powerful multi-core computing solutions such as GPUs.

The other dimension we focus on is the increasing sampling intensity leading to the availability of large numbers of sequences for viral evolutionary reconstructions. This is also illustrated by the sequencing efforts during the 2013-2016 West African Ebola virus outbreak that yielded over 1 600 complete genomes (Figure 1), representing over 5% of the known cases and making it the most densely sampled acute viral outbreak to date. This upscaling in sequencing is also impacting the molecular epidemiology of viruses with older transmission histories. For example, the genome sequencing efforts of the "Phylogenetics and Networks for Generalised HIV Epidemics in Africa" consortium (PANGEA-HIV) recently presented almost 4 000 HIV consensus sequences from different cohorts across sub-Saharan Africa (Ratmann et al., 2017). Such initiatives are transforming HIV molecular epidemiology into "big data" science, which hopefully can lead to new insights into HIV-1 transmission dynamics that can be translated to prevention strategies. However, large numbers of sequences also represent

a tremendous challenge for computational analyses, even more so than the length of the sequences. Indeed, it is well-known that phylogenetic likelihood computations scale linearly with alignment length, but the number of possible tree topologies grows super-exponentially with the number of taxa, which makes the search for optimal trees or distributions a highly cumbersome task. For this reason, we dedicate a large section in this chapter on phylogenetic approaches aimed at tackling the large data problem in viral sequence analyses.

Viral genomic data analyses are now frequently performed in the context of phylodynamics, a term that was originally introduced to describe "the melding of immunodynamics, epidemiology and evolutionary biology" (Grenfell et al., 2004). A key focus of phylodynamics is how the genetic diversity of viral pathogens is shaped by epidemic processes and natural selection (mostly in the context of immunological processes). Due to high mutation rates, large population sizes and short generation times, RNA viruses evolve at high evolutionary rates ensuring that their genomes can accumulate substitutions even over short-term epidemic time-scales. Sampling viral genomes over the time-scale of such epidemic processes therefore allows us to capture the relationship between time elapsed and sequence divergence. This is illustrated by plotting the root-to-tip divergence for the taxa in Ebola trees (Figure 1) as a function of sampling time in Figure 2, which can easily be done using software packages such as TempEst (Rambaut et al., 2016a). In this case, complete genomes show an increasing divergence over the sampling time range (Figure 2, left) while no such temporal signal is apparent in the corresponding glycoprotein gene sequences (Figure 2, right).

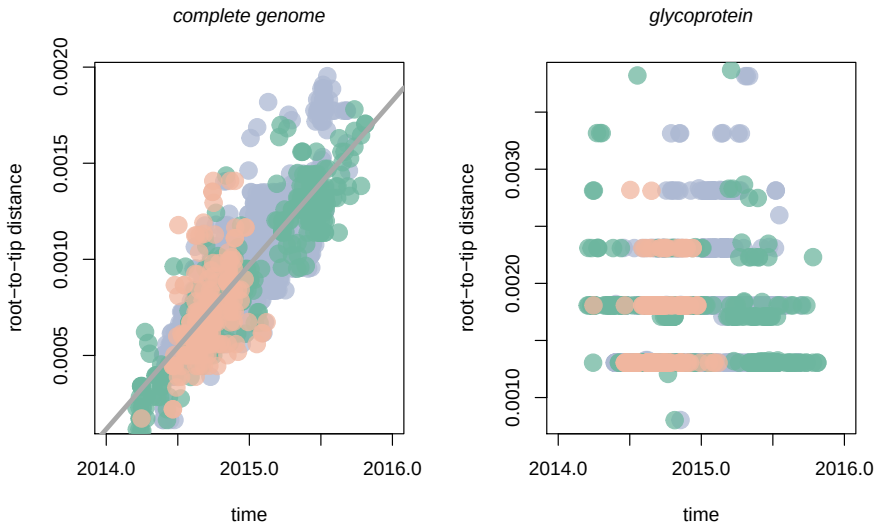


Figure 2 Plotting root-to-tip divergence as a function of sampling for the 2013 – 2016 Ebola virus data set. Data points are coloured according to country of sampling (green: Guinea; blue: Sierra Leone; red: Liberia). A comparison between complete genomes (left) and the glycoprotein gene (right) reveals a clear increase in divergence over the sampling time versus no temporal signal respectively.

This relationship between time and sequence divergence can be used to inform or calibrate molecular clock models (see Chapters 4.4 and 5.1 [Bromham 2020; Pett and Heath 2020]). These models are now routinely applied to phylogenetic trees, or integrated in phylogenetic inference, in order to estimate trees in units of time allowing us to date the epidemic origins or key (transmission) events in epidemic histories. Time-measured trees are also the necessary

prerequisite for the application of coalescent models that aim at estimating changes in epidemic size through time, and of birth-death models that estimate key parameters that determine an epidemic, such as the basic reproductive rate (R_0), i.e. the average number of cases one case generates over the course of their infectious period assuming a fully susceptible population (e.g. Fraser et al. 2009).

More broadly, pathogen genomes have been shown to contain a wealth of information concerning population size dynamics and the process of spatial spread that generated the geographic distribution of an epidemic, which can be recovered using a wide variety of demographic and phylogeographic models within a phylodynamics framework (Minin et al., 2008; Lemey et al., 2009, 2010; Gill et al., 2013, 2016). We illustrate some of the insights these approaches can obtain using specific viral examples.

Data

In this chapter we illustrate computational proaches on a data set of dengue virus (DENV), the most common vector-borne viral disease of humans. The dengue viruses are members of the genus *Flavivirus* in the family *Flaviviridae*, with four serotypes of dengue virus having been discovered to date. The four dengue serotypes are relatively closely related, but even within a single serotype, there is considerable genetic variation (Blok, 1985), with each serotype being further divided into several genotypes. The serotypes diverged approximately $1\,000 - 2\,000$ years ago; the genotypes within each serotype diverged much more recently, about $100 - 200$ years ago (Pollett et al., 2018). Despite these variations, DENV infections result in similar disease and clinical symptoms irrespective of serotype/genotype. Dengue serotypes are increasingly co-circulating in most regions of the world, particularly in Latin America and Asia (Messina et al., 2014), with global phenomena such as urbanisation and international travel acting as key factors in facilitating the spread of dengue.

We here perform phylodynamic inference on a dengue virus data set consisting of 997 genomes spanning the global dengue diversity, with a total of 6 869 unique site patterns across 10 gene-based partitions. This data set was generated in 2014 by downloading from GenBank all 3 289 available dengue genomes with known sampling year and country. Based on a maximum likelihood tree reconstruction, we used Phylogenetic Diversity Analyzer (Chernomor et al., 2015) to select the most diverse subset of 1 000 genomes. Three outliers according to root-to-tip regression explorations were excluded (Rambaut et al., 2016a). The nucleotide partitions correspond to the ten protein-coding genes that make up the dengue genome.

The 997 samples of this data set are annotated with discrete location states, one of the most popular traits associated with virus data sequences. This allows us to perform spatial ancestral state reconstruction in order to determine the origin of specific outbreaks and track viral spread over time. Owing to the widespread nature of dengue viruses, a total of 64 countries across six continents are used as location states to perform such a reconstruction. The number of taxa and the state dimensionality make for a computationally demanding joint inference of nucleotide and trait evolutionary processes.

In this chapter, we show how to analyse such a challenging data set using two popular inference frameworks: an approach oriented towards maximum likelihood inference (Section 2) and a fully Bayesian inference approach through Markov chain Monte Carlo (MCMC) (Section 3).

To illustrate the ability of maximum likelihood methods to handle very large data sets, we used an additional, larger, data set of 5 132 full dengue genomes downloaded from GenBank (Benson et al., 2012). The sequences were serotyped and genotyped with GenomeDetective (Vilsker et al., 2019) and those with type support < 100 removed. When

available, the sequences were annotated with the collection date and country metadata from the Entrez molecular biology database system (Sayers et al., 2009). The larger data set includes all the sequences from the smaller one, represents more (83 vs 64) countries, and is more noisy: no temporal outlier filtering or diversity-based selection was performed; moreover, some of the sequences had no sampling date (4%) or no location (1%) metadata.

2 Analysis of large phylogenies with Maximum Likelihood

Maximum likelihood (ML) inference is a procedure that for a given model finds the parameter values that maximise the observed data likelihood, thereby producing a single (maximum-likelihood) estimate of – for example – a phylogeny. Typically, an ML approach splits the analysis into several steps forming a pipeline, where the phylogeny reconstruction from sequence data is followed by further analysis (e.g. divergence time estimation, ancestral location reconstruction, ...) assuming a fixed phylogenetic tree. Using the dengue data set analysis as an example, we describe an ML pipeline for phylogeographic analysis of virus spread over time, consisting of the following steps detailed below: phylogenetic tree reconstruction from multiple sequence alignment, tree rooting and dating, and ancestral character reconstruction for geographic data.

2.1 Tree reconstruction

Phylogenetic inference using ML aims at finding the tree and model parameters that maximise the likelihood and is known to be NP-hard (Chor and Tuller, 2005). ML tree reconstruction tools generally approach the problem by performing a "hill-climbing" optimisation, i.e. these methods start by generating an initial tree (e.g., a randomly generated tree or a maximum-parsimony tree), and then keep replacing it with a better (in terms of likelihood) neighbouring tree (obtained using certain topological rearrangements), until no better tree can be found. A potential danger in a pure hill-climbing is the possibility to get trapped in a local optimum. To overcome this issue, it is recommended to perform the optimisation starting from multiple initial trees, and then keep the best result, and to use more expansive techniques for searching neighbouring tree space (Zhou et al., 2018). The most common topological rearrangement algorithms for finding a neighbour tree are Nearest-Neighbour-Interchange (NNI), which swaps two non-sibling subtrees adjacent to an internal branch (Robinson, 1971), and Subtree-Pruning-and-Regrafting (SPR), which prunes a subtree from the initial tree and regrafts it onto a different branch (Swofford et al., 1996). SPR can evaluate many more trees (quadratic to the number of tips) from one initial topology than NNI (linear), but as a consequence it is also much slower (Allen and Steel, 2001) and therefore different heuristics have been developed to filter out the unpromising SPR candidates (Stamatakis et al., 2005; Hordijk and Gascuel, 2005; Guindon et al., 2010).

Among multiple ML tools that are available for phylogenetic tree reconstruction, the most popular are FastTree (Price et al., 2010), PhyML (Guindon et al., 2010), RAxML (Stamatakis, 2014) (and its updated version RAxML-NG [Kozlov et al. 2019], see also Chapter 1.3 [Kozlov and Stamatakis 2020]), and IQ-TREE (Nguyen et al., 2015). FastTree works very well to perform a preliminary analysis as it can be orders of magnitude faster than other ML tools, but as a trade-off, generates less accurate tree estimates due to limited tree space exploration and less thorough branch length optimisation. FastTree starts with a distance-based optimisation using both NNI and SPR rearrangements, followed by ML-based NNI rearrangements to search for the final tree, using heuristics at all stages to limit the numbers of tree searches and likelihood optimisations. PhyML performs hill-climbing tree searches using both NNI and

SPR rearrangements (with a parsimony-based filtering of the least promising SPR moves), while RAxML/RAxML-NG implements SPR-based hill-climbing, employing heuristics to reduce the number of unpromising SPR candidates (typically regrafting the pruned subtree in a position remote from the original one). IQ-TREE combines hill-climbing algorithms with random perturbations of the current best trees and broad sampling of initial starting trees, and generates a pool of candidates containing the top 5 trees obtained by NNI hill-climbing from the 20 best parsimony-based starting trees. At each iteration, a randomly chosen candidate tree is perturbed with $0.5(n-3)$ random NNIs, where $n-3$ is the number of inner branches, and optimised with the hill-climbing NNIs. If the resulting tree is better than the best tree in the candidate pool, the iteration is considered successful and the best tree is replaced. Otherwise the worst tree in the candidate pool is replaced if it is worse than the resulting tree. The analysis terminates after a certain number of unsuccessful iterations. RAxML (and especially RAxML-NG) and IQ-TREE are parallelised which makes them faster than PhyML.

In a comparison of these different ML packages in terms of their accuracy, Zhou et al. (2018) recently analysed various middle-sized data sets of about 200 taxa, including genome and transcriptome data of fungi, animals and plants, using both single gene alignments as well as concatenated full genomes. In this comparison, IQ-TREE (version 1.5.5) outperformed RAxML (8.2.11) and PhyML (20160530) in most cases, with the exception of some of the largest data sets. Kozlov et al. (2019) repeated the comparisons on the same data sets using the latest available version of RAxML (RAxML-NG) and found it to be generating the highest tree likelihood while being 1.3 to 4.5 times faster than IQ-TREE. Similar results can be obtained with the most recent (but as of yet unpublished) version of PhyML (`https://github.com/stephaneguindon/phyml`; data not shown), showing that there is still room for improvement in these complex algorithms and programs.

In conclusion, for extremely large trees (dozens of thousands of tips) FastTree is probably the only choice due to its speed, while for the other cases (up to thousands of tips) we recommend to use several programs and compare the results.

2.1.1 Application to dengue data

The four dengue serotypes diverged thousands of years ago while the genotypes within each serotype diverged about 100 to 200 years ago (Pollett et al., 2018). We therefore expect a phylogeny with very long branches connecting the serotypes and much shorter ones within each serotype. This makes the common tree reconstruction challenging: On one hand, for deep phylogenies (like the inter-serotype one) one often uses amino acid alignments due to codon degeneracy and therefore loss of signal for the deep nodes on the nucleotide level (Rota-Stabelli et al., 2013). On the other hand, for the intra-serotype phylogenies, which contain much closer related sequences, amino acid alignments might not have enough resolution, and hence the nucleotide alignments should be preferred. To account for codon degeneracy a partitioning scheme with a distinct group for the third codon positions can be used.

We have reconstructed phylogenies for the smaller (997 sequences) and the larger (5 132 sequences) data sets with RAxML-NG (version 0.9.0, starting from a parsimonious tree) and IQ-TREE (version 1.6.9), both from the amino acid alignment (HIVb+I+G6 evolutionary model) and from the nucleotide one (GTR+I+G6) with a two-group partitioning scheme (Chernomor et al., 2016): for the codon positions $1-2$ and for the position 3. Evolutionary models were selected by the model selection tool SMS (Lefort et al., 2017), we increased the number of GAMMA categories from 4 (as used in SMS) to 6, for a better fit of

the gamma distribution of rates across sites.

On a machine with 12 cores, phylogeny reconstruction for the larger data set took 1 day 8 hours for RAxML-NG on the amino acid alignment, and 7 hours on the nucleotide alignment; for IQ-TREE it took 1 day 2 hours on the amino acid alignment, and 1 day 17 hours on the nucleotide alignment. In terms of likelihood RAxML-NG got a better result on the amino acid data (log likelihood of −174 295 vs −174 473), and IQ-TREE on the nucleotide one (−873 396 vs −873 406).

Once the reconstruction is done it is important to assess the results. It can be done using prior knowledge on the expected tree topology, and by comparing the phylogenies obtained with different tools. The tree topologies of the reconstructed phylogenies were overall close, especially those reconstructed on the same alignment. We formally assessed this using the normalised quartet distance (Estabrook et al., 1985) (calculated with tqDist [Sand et al. 2014]), which takes on values between 0.0 for identical trees and 1.0 for trees that have no quartet in common. The normalised quartet distance was 0.004 [DNA] and 0.008 [AA] for the tree pairs reconstructed with different tools on the same alignment, and varied from 0.015 (RAxML-NG [DNA] vs RAxML-NG [AA]) to 0.022 (IQ-TREE [DNA] vs IQ-TREE [AA]) for the tree pairs reconstructed on different alignments.

However, the reconstructed topologies (for both data sets) had differences for the deep parts of the phylogeny, i.e. serotype subtree rooting. Moreover, none of them corresponded to the the expected tree topology (prior knowledge). We expected to find monophyletic clades for the genotypes within the same serotype with a serotype root that is placed on one of the inter-genotype branches as done by Pollett et al. (2018); we will also confirm this assumption in Section 2.2 with rooting of serotype-specific trees based on dates. For DENV2 and DENV4 this was the case: the root was placed on the branch separating sylvatic genotype from the epidemic ones, all the genotypes were monophyletic, and their relative positions were consistent in all the reconstructed phylogenies. For DENV1 and DENV3 the relative genotype positions were consistent in all the phylogenies, but the root positions varied and the root was often placed within one of the genotypes (see Figure 3). The only exception was the phylogeny reconstructed by IQ-TREE on the nucleotide alignment: it managed to place the DENV1 root correctly (but still had issues with DENV3).

This difficulty with resolving deep parts of the phylogeny is likely due to the fact that all the DENV1 and DENV3 genotypes diverged relatively simultaneously (within dozens of years, while the root age is $1\,000 - 2\,000$ years), and there is a lack of sequences that diverged earlier (e.g. within hundreds of years). Moreover, the branches connecting serotype subtrees were extremely long: $0.56 - 1.1$ mutations per site. Bayesian methods can address this issue by incorporating temporal and phylogeographic information along with the alignment data, which increases the signal. In the maximum clade credibility (MCC) tree reconstructed with Bayesian methods (see Section 3), the rooting of DENV1, DENV2 and DENV4 is correct, however DENV3 subtree root is also misplaced within one of the genotypes (3.II).

To overcome the issues described above, we reconstructed serotype-specific subtrees from the nucleotide alignments using both RAxML-NG and IQ-TREE (as described before) and then kept the most likely tree for each serotype.

2.2 Tree rooting and dating

For data sets in which sufficient genetic change has accumulated during the window of sampling, the divergence in each sequence is expected to correlate with the date of sampling, as illustrated in the introduction (Figure 2). Hence, the sampling times of the sequences can be used to estimate the substitution rate and the divergence dates, transforming a phylogeny

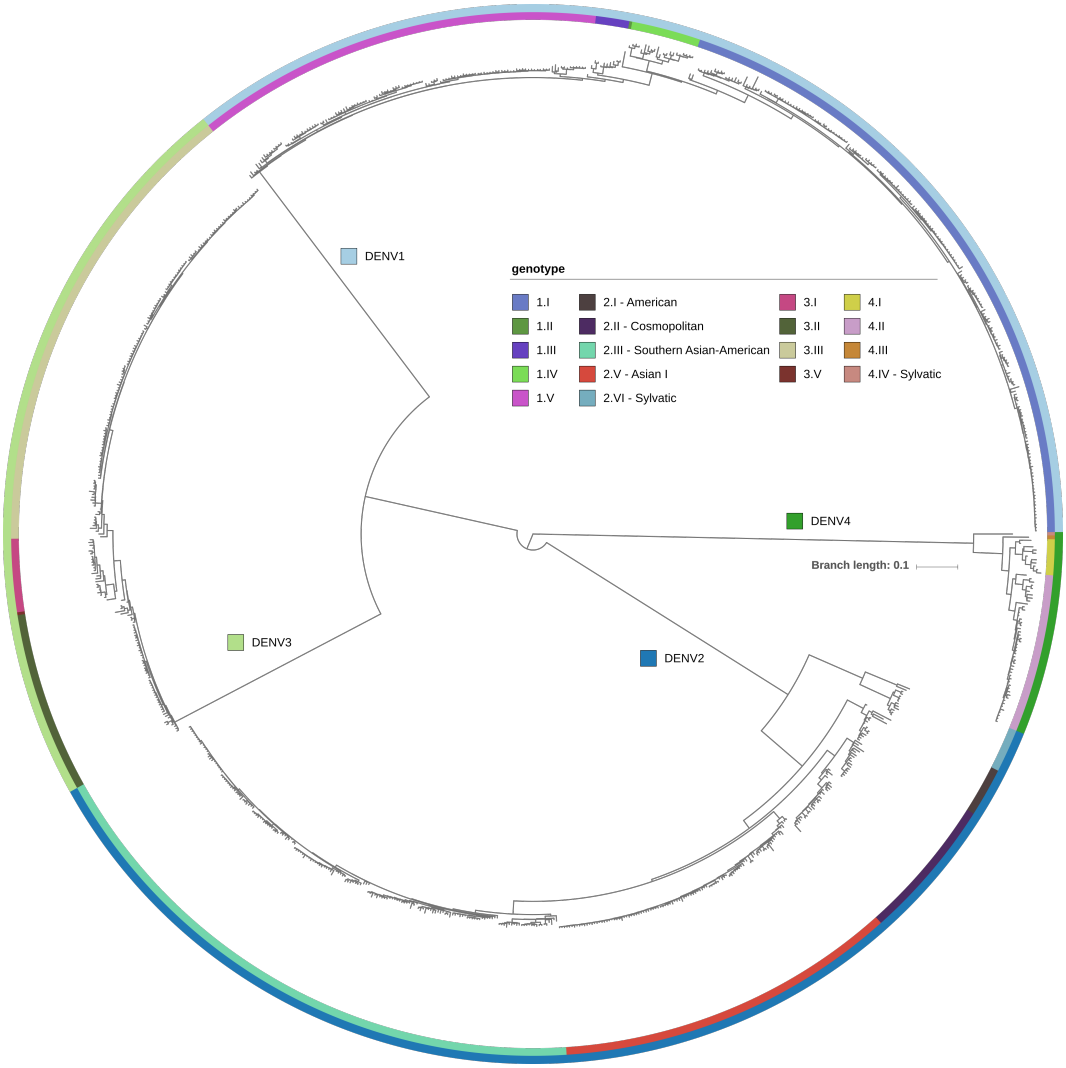


Figure 3 Phylogeny estimated with RAxML-NG on the nucleotide alignment for the small data set, visualised with iTOL (Letunic and Bork, 2007). The rooting of the DENV2 and DENV4 serotype subtrees is correct, with all the genotypes being monophyletic and the roots placed on the branches separating sylvatic genotypes from the epidemic ones. For DENV1 and DENV3 subtrees the roots are misplaced: instead of a branch separating different genotypes, the serotype roots are within one of their genotype trees (3.II for DENV3 and 1.V for DENV1), however the other genotypes are monophyletic.

into a time-scaled tree whose branches are measured in units of time, e.g. years. Various molecular clock models are available to perform such estimations, such as the strict clock (SC) – which assumes substitution rate homogeneity throughout the tree (Zuckerkandl and Pauling, 1965) – and the uncorrelated relaxed clock – which allows a different rate along each branch in the tree, but with rates assumed to follow a chosen statistical distribution (Bromham et al. 2018; Chapter 4.4 [Bromham 2020]). The methods for substitution rate and time-scaled tree estimation can be divided into two groups: those that incorporate sequence dates into the tree reconstruction framework, and the ones that estimate the substitution rate on a fixed phylogeny (with fixed branch lengths measured in substitutions per site).

The methods of the former type evolved from an intermediate approach where only the tree topology was fixed (but not the branch lengths) and were initially implemented in the program TipDate (Rambaut, 2000) (and later as an R package node.dating [Jones and Poon 2017]): given a rooted tree topology and the tip dates, TipDate optimises the substitution rate under the SC model and estimates the times of the internal nodes using ML under a given evolutionary model, e.g. HKY (Hasegawa et al., 1985). Drummond et al. (2001) first extended TipDate by allowing different rates to be estimated for different intervals of time, and later embedded them into a Bayesian framework for joint inference of mutation rate and population size that incorporates the uncertainty in the genealogy by using MCMC integration (Drummond et al., 2002).

Among the methods of the latter type, one of the very first ones was Root-To-Tip (RTT) regression (Shankarappa et al., 1999; Drummond et al., 2003a): assuming a strict clock, the root-to-tip distance in the phylogeny should be proportional to the corresponding elapsed time, and a regression of the root-to-tip distance as a function of tip dates provides estimates of the mean substitution rate (regression slope) and the root date (x-intercept). This constitutes a very fast method that allows estimating the root of the tree, e.g. by searching for a tree branch that minimises the sum of regression residues. However, it does not provide the dates for the internal nodes, and therefore does not output the time-scaled tree. Also, RTT regression violates the assumption of data independence as deep branches contribute to multiple RTT distances, it therefore is not suitable for statistical hypothesis testing and should rather be used as a data exploration tool (Rambaut et al., 2016b).

The LF (Langley and Fitch, 1974) model – implemented in r8s (Sanderson, 2003) – assumes a strict clock with a constant substitution rate, and a Poisson distribution for the number of substitutions along every tree branch. The substitution rate and the internal node dates are estimated by maximising the likelihood of the rooted input tree.

Least-Squares Dating (LSD) (To et al., 2016) uses a normal approximation to the LF model to estimate the substitution rate. LSD assumes the following relationship between the branch lengths in the initial phylogeny and the corresponding time-scaled tree:

$$b_i = y_i \cdot \omega + \epsilon_i,$$

where b_i is the length of the branch i measured in substitutions per site, y_i – its length in years, ω is the substitution rate, and $\epsilon_i \in N(0, \sigma_i^2)$ is a noise (error) term drawn from a normal distribution (independent for different branches). In other words, LSD assumes a strict clock, but the noise term makes it robust to uncorrelated violations. When run in "temporal precedence constrained mode", LSD additionally ensures that all the dated branch lengths are non-negative (which could otherwise be violated due to the noise terms for short branches). LSD minimises the error using the weighted least squares criterion:

$$\sum_i \frac{1}{\sigma_i^2}(b_i - y_i \cdot \omega) \rightarrow min,$$

where the variance terms are derived from the Poisson nature of the substitution process as

$$\hat{\sigma_i^2} = \frac{b_i + c/S}{S},$$

S being the sequence length, and c – a constant smoothing factor. Uncertainty can be obtained via parametric bootstrap or by repeating the analyses on a set of non parametric bootstrap trees. LSD can root and date large phylogenies in quadratic time (i.e. proportional to the number of tips squared). The new (but as of yet unpublished) version of LSD – LSD2 (`https://github.com/tothuhien/lsd2`) – extends the tool with several new features, including outlier detection (sequence or date annotation errors or samples that are poorly described by the fitted substitution model), and the local clock model (Yoder and Yang, 2000).

Treedater (Volz and Frost, 2017) extends the concepts of LSD by implementing outlier detection and an uncorrelated relaxed molecular clock (i.e. the global substitution rate ω is replaced with a collection of branch-specific rates ω_i drawn from a common gamma distribution). Treedater implements a heuristic iterative approach to optimise both the parameters of the gamma distribution (shape and scale) and the internal node times. It initialises branch-specific rates to the common rate estimated by RTT, then repeats the optimisation cycle until convergence (defined by a tolerance threshold). At each iteration, first the internal node times are calculated based on the branch-specific rates by solving a (constrained) least-squares problem like in LSD, secondly the gamma distribution parameters are optimised based on the new ancestral dates using gradient-descent. Therefore when compared to LSD, the computation time is multiplied by the number of iterations. Moreover, it is recommended to repeat the optimisation with different starting conditions to avoid local optima.

Treedater also implements a statistical test for selecting the appropriate clock model. In a comparison performed by Volz and Frost (2017), treedater outperformed LSD and BEAST on 9 out of 16 simulated data sets, with all methods showcasing their strengths and weaknesses in at least one scenario.

TreeTime (Sagulenko et al., 2018) is an ML relaxed clock method that allows to either optimise the branch lengths on a given tree topology (along with ancestral sequence reconstruction) or to use the branch lengths provided in an input phylogeny. TreeTime uses dynamic programming for branch length optimisation and its run times scale linearly in the size of the data set.

Duchêne et al. (2016b) have compared the rate estimation by Bayesian, least-squares and RTT regression methods on 81 RNA and DNA virus data sets (9 to 120 sequences of 350 to 10 066 nucleotides sampled over 0.5 to 86 years), and observed that the methods largely produce congruent estimates of substitution rates, provided that the data meet certain criteria, such as the absence of high among-lineage rate variation, congruence between the tree topology and no phylogenetic and temporal clustering. Moreover Duchêne et al. (2016b) pointed out that clock-model testing should be routinely performed, as the use of relaxed molecular clocks can lead to overestimates of the mean substitution rate when the data in fact fit a strict clock.

2.2.1 Application to dengue data

ML tree reconstruction methods estimate an unrooted phylogeny, which one can either root using an outgroup, or using the sampling dates and a molecular clock model. In the case of the dengue data set, no outgroup is present for the full tree (though different serotypes/genotypes

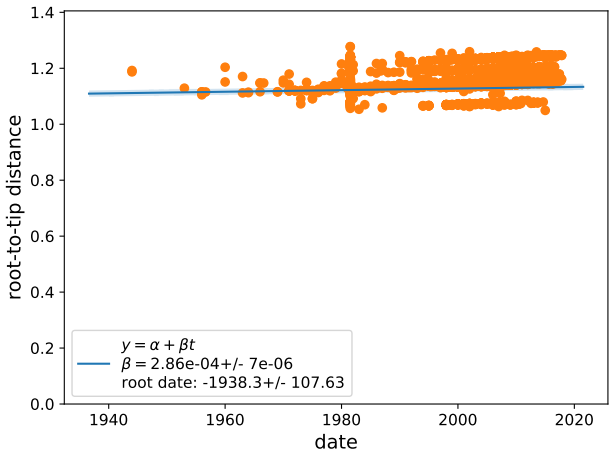


Figure 4 Results of RTT regression (performed using TreeTime) for the large dengue data set (RTT regression on the small data set is similar). The RTT distance is plotted as a function of the tip dates and provides estimates of the mean substitution rate (regression slope) and the root date (x-intercept). The fit line is almost horizontal, therefore an accurate root date estimation is almost impossible: the slightest error in the slope would lead to a large error in the x-intercept.

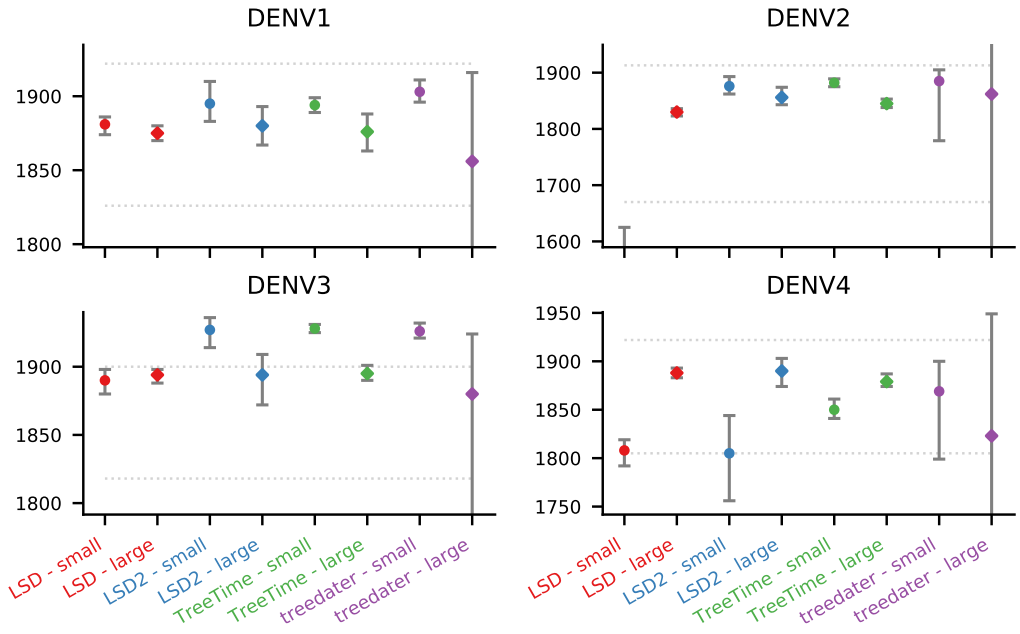


Figure 5 Estimated dates of emergence of each of the four dengue human serotypes [years] (with 95% CIs), obtained with treedater, TreeTime, LSD, and LSD2. The estimates obtained on the smaller data set are shown with circles, on the larger one – with diamonds. Median 95% CIs from the literature that are summarised in Table 3 of (Pollett et al., 2018) are shown by dotted horizontal lines. The estimates provided by different tools are generally within the CIs from the literature (apart from DENV3 small data set and LSD estimate for DENV2 small data set (due to outliers)) but vary for different data sets and tools, which could be due to not enough quality control for the samples included in the data sets.

can serve as outgroups for its subtrees), and we therefore had to resort to using sampling dates for rooting. RTT regression suggests the presence of temporal signal in the data, but potentially too short of a sampling window (with respect to the total time span from the most recent common ancestor up to the present) for the full tree: the x-intercept (root date) of the RTT plot for the full tree (Figure 4) is very old, while the slope (rate) is almost horizontal, therefore the slightest error in the rate estimation (slope) would result in a large error in the root date (x-intercept). Moreover, underestimation of evolutionary distances in deep branches separating serotypes might obfuscate temporal signal (Duchêne et al., 2016a). If we had ancient samples, their dates might have helped to calibrate the slope. Dating, rooting and confidence interval (CI) estimation (the most time consuming part) on the full tree was performed in 2 minutes by LSD/LSD2, in 1 hour 30 minutes by TreeTime, and in 6 days by treedater.

To assess the performance of various ML packages on more recent data, we dated the four serotype phylogenies separately. We used four recently developed applications that allow dating unrooted phylogenies: LSD (version 0.3beta), LSD2 (version 1.4, strict clock), treedater (version 89a0df0) and TreeTime (version 0.5.5, using branch lengths of the input tree). TreeTime and treedater implement both relaxed and SC, and are able to detect outliers, while LSD only implements SC and does not detect outliers. LSD2 extends LSD with outlier detection.

The clock model selection test performed by treedater identified the relaxed clock as the model of choice, therefore we run treedater and TreeTime with the uncorrelated relaxed clock model. For rooting we used the sylvatic outgroup for DENV2 and DENV4, and estimated the root position from dates for DENV1 and DENV3. Note that not all of the outliers detected by different methods were the same. TreeTime generally finds more outliers than LSD2 which in turn finds more than treedater, e.g. for DENV4 subtree TreeTime detected 35 outliers, LSD2 detected 14 (5 of which were among the ones detected by TreeTime), while treedater detected none.

The rates estimated by different tools were close to those reported in the literature (see Table 2 in (Pollett et al., 2018) for a summary) for DENV1-2 and about 2 times lower for DENV3-4. Figure 5 shows the estimated dates of emergence of the four serotypes, where the dotted horizontal lines indicate the median 95% CIs from a literature summary provided in Table 3 of (Pollett et al., 2018). All the estimates are consistent with literature, except for those obtained on DENV3 small data set (potentially due to root position as explained below) and the LSD estimate on DENV2 small data set (due to outliers). Additionally, we observe several phenomena: (1) As expected, the full tree was harder to date than the serotype trees, which led to much larger CIs for the dates, e.g. $[-6392, -1957]$ for treedater (data not shown); (2) The estimates provided by LSD2 (after outlier removal) are often different from those of LSD (with outliers), suggesting an important impact of outliers; (3) CIs calculated by treedater are much larger and generally include CIs from other methods. CIs estimated by different tools on the large data set overlap more, suggesting that adding more data helps to increase the signal; (4) The estimates provided by different tools are quite consistent for DENV1 but much less so for other serotypes, which could be due to the absence of quality control for the samples included in the data sets, such as removal of clone sequences, duplicates, erroneous metadata, etc.

The estimated root position was consistent among all tools and data sets for the full tree (on the branch separating DENV4 from the other serotypes) and DENV1 tree (on the branch separating the common ancestor of genotypes III and V from the other genotypes); and varied for DENV3 tree. For DENV3 three scenarios for the root position were present: (1)

on the branch separating genotype II from the other genotypes (LSD, LSD2 and TreeTime on the small data set); (2) on the branch separating the common ancestor of genotypes II and III from the common ancestor of genotypes I and V (LSD, LSD2 and TreeTime on the large data set, and treedater on the small data set); and (3) unresolved root, with genotype II, genotype III and the ancestor of genotypes I and V as children (treedater on the large data set). Note that (3) represents a consensus between (1) and (2), moreover the branches that needed to be collapsed to reach this consensus with the time-scaled trees obtained by other tools were short (1.5-5 years).

Overall, our analyses indicate that dating an unrooted tree is a complex task, especially when combined with using a relaxed clock and outlier detection and removal. In theory, the use of a relaxed clock should reduce the number of outliers compared to a strict clock but both depend highly on the root position. When applied to noisy (real) data, dating becomes truly challenging.

2.3 Phylogeography

Analyses of epidemic spread through space and time are often performed by defining a finite, non-ordered set of locations (e.g. countries) and using ancestral character reconstruction (ACR) along a phylogenetic tree with the defined locations as possible character states, based on the known tip locations. ACR aims to unravel how the character has changed on the tree from the root to the tips through time, by assigning the most likely ancestral character state to each internal node. Various inference methods can be used to perform ACR, and a range of software packages is available for each type of inference.

Parsimony-based ACR methods infer a scenario with minimum state changes along the tree. These methods are quick and simple, however, due to the over-simplification of evolutionary processes (e.g. not accounting for branch lengths and evolutionary times), parsimony has limited accuracy (Zhang and Nei, 1997; Collins et al., 1994). ML and Bayesian approaches on the other hand are based on probabilistic models of character evolution that adapt standard nucleotide substitution models to s-state (discrete) trait characters. They have been shown to outperform parsimony methods, using both theoretical arguments and simulation studies under a variety of conditions (Zhang and Nei, 1997; Gascuel and Steel, 2014), and are also robust to moderate model violations and phylogenetic uncertainty (Hanson-Smith et al., 2010).

Statistical ACR methods employ continuous time Markov chains (CTMCs) models that emit discrete outcomes as a continuous function of time. This process is assumed to be memoryless, in that the probability of transitioning to a new location only depends on the current location and not the past history. As with nucleotide substitution models, an $s \times s$ infinitesimal rate matrix $\Lambda = \{\lambda_{ij}\}$ completely characterises the CTMC process (Lemey et al., 2009). The rate matrix Λ contains non-negative off-diagonal entries and all rows sum to 0, yielding a stochastic matrix upon exponentiation. To determine the finite-time transition probabilities between states (or locations) over a branch of length t, the following matrix exponentiation is required:

$$P(t) = e^{\Lambda t} \tag{1}$$

In its most general form, such a discrete trait substitution model allows a unique instantaneous rate between each pair of character states to be estimated, which may prove problematic because these rate parameters are only informed by a single discrete trait character observed

at the tips of the phylogeny (Gascuel and Steel, 2019). Hence, simpler models are often used, such as s-state generalisations of 4-state JC and F81 models for DNA (Jukes and Cantor, 1969; Felsenstein, 1981). Under F81-like models, migration rate from a state (location) i to a state j ($i \neq j$) is proportional to the equilibrium frequency of j, π_j; JC-like models are a special case where all equilibrium frequencies are equal: $\forall i\ \pi_i = 1/s$. A big computational advantage of F81-like models is that the probability of changes along a branch of length t can be calculated with a simple formula:

$$P_{i\to j}(t) = \left\{ \begin{array}{ll} (1-e^{-\mu t})\pi_j, & \text{if } j \neq i \\ e^{-\mu t} + (1-e^{-\mu t})\pi_j, & \text{otherwise} \end{array} \right\}, \text{where } \mu = 1/(1-\sum_i \pi_i^2). \quad (2)$$

Dudas et al. (2017) showed that the origin and destination population sizes (π_i and π_j) are two of the main factors explaining Ebola dissemination in West-Africa. This advocates for the use of s-state F81-like models, where the expected number of changes from i to j is proportional to $\pi_i\pi_j$. Moreover, Gascuel and Steel (2014) showed with simulations on DNA-like data generated using an HKY model (Hasegawa et al., 1985) with high transition/transversion rate and heterogeneous nucleotide frequencies that even the simpler JC-like model performs nearly as well as the true one.

In the likelihood framework, to predict ancestral character states based on the selected model of character evolution, one commonly uses the marginal posterior probabilities of the character states (Felsenstein, 1981; Yang, 2007), the joint reconstruction of the most likely scenario (Pupko et al., 2000), or an approach that lies in between these two extremes. Marginal reconstructions provide users with state probabilities, but these are difficult to interpret and visualise, while joint reconstructions select a unique state for every tree node and thus do not reflect the uncertainty of inferences. Intermediate approaches overcome these limitations by predicting a unique state in the regions of the tree that are easy to estimate (typically close to the tips [Gascuel and Steel 2014]), while keeping several likely states in the more difficult regions (typically close to the root), reflecting the uncertainty of the inferences.

PastML (Ishikawa et al., 2019) is a fast ACR tool that implements several parsimony and ML ACR methods: Joint, MAP (maximum a posteriori) that selects the state with the highest marginal probability, and MMPA (marginal posterior probabilities approximation), an intermediate approach that uses decision-theory concepts and the Brier criterion to associate each node in the tree to a set of likely states: a unique state is predicted in the tree regions with low uncertainty, while several states are predicted in the uncertain regions.

An important choice to take before performing a phylogeographic ACR is that of the precision level and whether the geographic sampling captures the range of the pathogen. Choosing a character with many possible states on a small data set can be limiting in terms of phylogeographic signal and can leave many nodes unresolved between several states for methods like MPPA, or predict a poorly supported state (e.g. low marginal probability even for the most likely state) for unique-state methods like Joint or MAP. Moreover, if the states present in the data (tree tips) do not cover all the possibilities, it can bias the predictions, as for instance it is impossible to predict missing countries. Biases in sampling for the states that are represented can also affect the reconstructions. Such biases are prominent in most real data sets, including the examples we study here.

To strengthen the signal extracted from the data, one needs to increase the number of sequences for each character state, sampled over a larger time range. This can be achieved

either by adding more sequences annotated with each state to the data set, or by generalising the annotations to decrease the number of character states, by grouping countries into broader geographic regions.

2.3.1 Application to dengue data

We reconstructed ancestral geographic characters with PastML (version 1.9.20) using the MPPA method and the F81-like model. Performing the country ACR on the full tree of the large data set (4 767 tips after outlier removal by LSD2) took 3 hours 30 minutes. The large data set includes sequences from 83 different countries, of which 65 are present in the smaller data set. Moreover 23 countries are present only once and hence do not provide sufficient information on the migration to and from them. As a result, 3.7% of internal nodes remain unresolved between several countries (10.3% for the small data set). As expected, the majority of unresolved nodes are found deeper in the tree, corresponding to the root and the common ancestors of different serotypes and genotypes.

The difference in percentage of unresolved nodes between the small and large data sets shows that adding more data increases the signal. The ACRs for the two data sets are generally compatible, while some of the nodes that remain unresolved in the small data set are resolved in the large one (see Figure 7), suggesting that the method is to some extent robust against sampling variation.

To further increase the signal we generalised the countries into 11 geographic regions: South America, Caribbean, Central America, Northern America, Africa, Europe, Western Asia, Eastern Asia, South-eastern Asia, Southern Asia, and Oceania. This allowed to reconstruct states closer to the root, and reduce the percentage of unresolved internal nodes to 1.2%. Finally, when we generalised the locations even further, into five continents (Americas, Africa, Europe, Asia, Oceania), the number of unresolved internal nodes was reduced to only 0.5%.

Summarised ancestral scenarios for location reconstruction on the full tree and for country on the DENV3 subtree are shown in Figures 6 and 7. PastML visualises these scenarios by (1) clustering the parts of the tree where no state change happens into meta-nodes (whose size corresponds to the number of samples (tips) they contain); (2) clustering independent events of the same kind into meta-edges (whose size corresponds to the number of such events). For example, the large "South-eastern Asia 1 503" node in Figure 6 represents the reconstructed cluster of DENV1 spread in South-eastern Asia, which includes 1 503 sequences in our data set; while its "Eastern Asia $1-7$" child node connected by a meta-edge of size 36 represents 36 independent DENV1 transmissions from South-Eastern Asia to Eastern Asia.

The predicted ancestral locations generally agree with previous studies, performed on different dengue data sets. For instance, in the study of global DENV2 phylogeography by Walimbe et al. (2014) performed using Bayesian inference on 307 DENV2 E-gene sequences from GenBank (sampled between 1944 and 2011), the authors also could not pinpoint the ancestral location for sylvatic and epidemic strains, and detected Southeast Asia as the ancestral region for the Asian/Asian-American and Cosmopolitan genotypes. The authors also found multiple migrations from the Caribbean to the American mainland (see Figure 6 for our predictions).

In the study of spatio-temporal dynamics of dengue in Colombia, performed on 143 newly sequenced samples from Colombia (sampled between 1998 and 2015) combined with full-length E-gene sequences retrieved from GenBank, Jiménez-Silva et al. (2018) performed a Bayesian analysis of spatial spread and detected significant viral diffusion between Venezuela

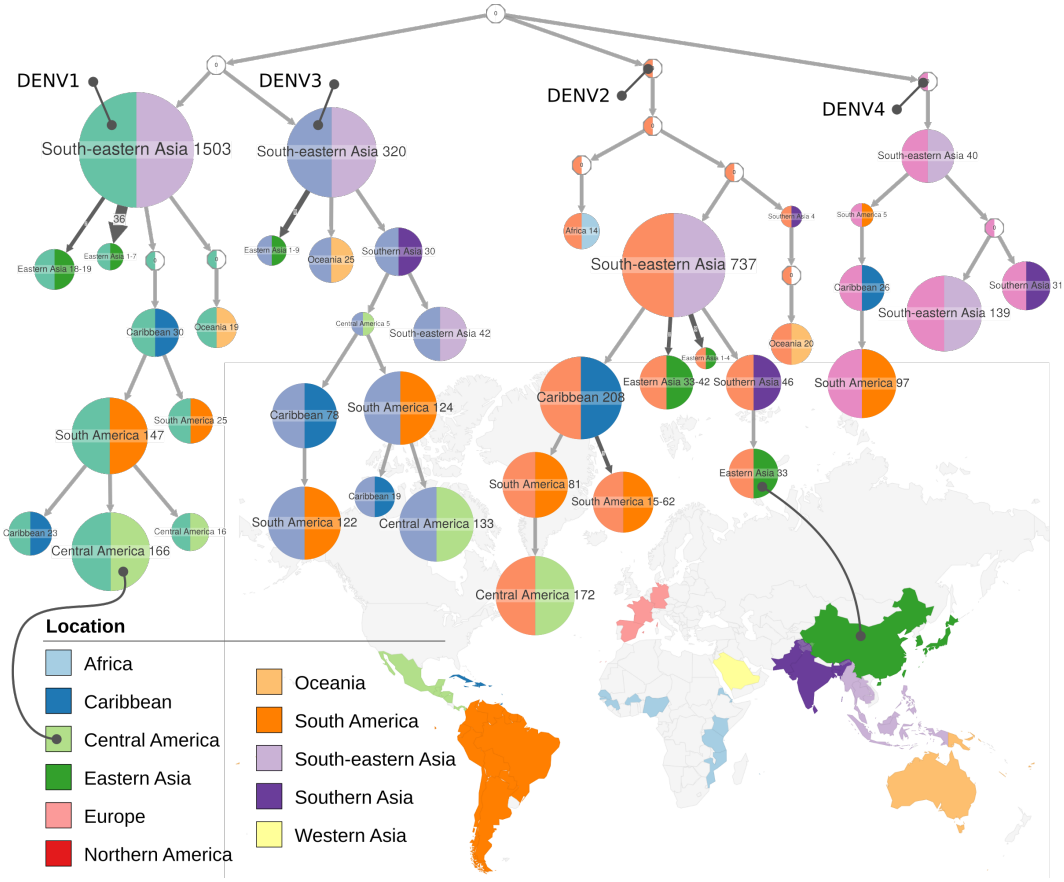


Figure 6 Ancestral location scenario reconstructed on the full DENV tree (large data set) using PastML (MPPA+F81). Locations are colour-coded in the right halves of the nodes and shown as labels, serotypes are colour-coded in the left halves of the nodes, as follows (from left to right): DENV1 (green), DENV3 (blue), DENV2 (orange), DENV4 (pink). Tips representing less than 14 sequences are not shown.

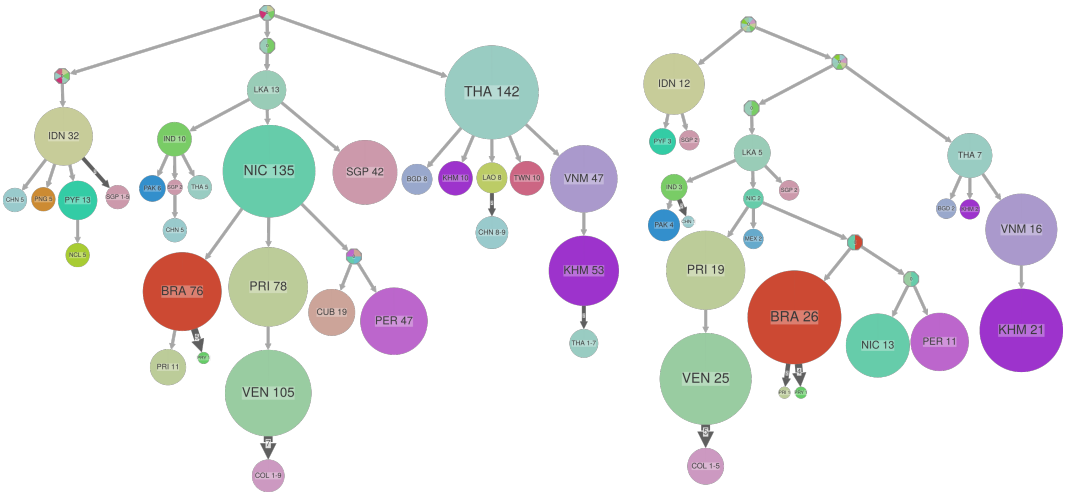


Figure 7 Ancestral country reconstructed on the DENV3 subtree with PastML (MPPA+F81) for the large data set (left) and the small one (right). Countries are colour-coded and shown as labels. Tips representing less than 5 (2) sequences are not shown for the large (small) data set. ACR for the two trees are compatible, additional sequences present in the larger data set allowed to resolve some of the nodes unresolved for the smaller one.

and Colombia for all serotypes. In our predictions using PastML we also observe this pattern, e.g. for DENV3 we inferred 7 introductions from Venezuela to Colombia (meta-edge of size 7 connecting "Ven 105" cluster to "COL $1-9$" one in Figure 7).

Tan et al. (2018) performed a Bayesian analysis of DENV3 genotype III spread to Malaysia using 602 complete coding sequences and 972 E-gene sequences from 56 countries between 1966 and 2014, including the complete genome sequences of 21 newly sequenced Malaysian DENV3 genotype III isolates. They detected an introduction from Sri Lanka to Singapore and from there to Malaysia. Our data set does not contain Malaysian sequences, but the spread of DENV3 genotype III from Sri Lanka to Singapore is also inferred, as indicated by the arrow from cluster "LKA 13" to "SGP 42" in Figure 7.

2.4 Discussion

ML methods are fast and permit the analysis of a large number of sequences in a relatively short time (e.g. 1.5 days for a phylogeny reconstruction on a 5 132 full genome data set with IQ-TREE, 2 minutes of dating with LSD2, and 3.5 hours of ancestral country reconstruction with PastML, on a 12-core machine). Adding more sequence data is desirable as it generally helps to reduce bias and uncertainty (e.g. for phylogeny rooting or ACR).

Phylogeny reconstruction is robust when applied to data with homogeneous time scale (e.g. serotype-specific trees): the results obtained by different tools were highly similar. However care needs to be taken when reconstructing phylogenies with mixed time scale (e.g. deep intra-serotype branches versus much shorter inter-serotype ones in the case of dengue): one needs to verify the reconstruction results, for example, based on prior knowledge on expected tree topology (e.g. monophyletic genotypes within the dengue serotypes), or by comparing the results obtained with different tree reconstruction tools. Rooting and dating with dengue data proved to be a difficult task. Popular ML/distance dating methods (e.g. LSD2, TreeTime) are fast and able to deal with large phylogenies, but particular problems remained difficult to solve (rooting, relaxed molecular clock, outliers, noisy dates). The full

dengue data set did not have enough signal for confident dating of deep nodes, and even for some of the more recent, serotype-specific phylogenies, different results were obtained depending on the method used. Any additional input (such as an outgroup) could be of great help. Preliminary exploratory analyses are very important, e.g. checking temporal structure and using regression to select an appropriate model and to determine whether molecular clock-based dating is valid at all. There is still room for improvement of these fast dating approaches.

ACR and phylogeography on the other hand, showed a strong and robust signal. A fast ML method like PastML was able to analyse large data sets, and provide results that were robust against sampling variations and phylogenetic uncertainty (see results for DENV3 [Figure 7] and [Ishikawa et al. 2019]).

3 Bayesian phylodynamic inference

Within the field of pathogen phylodynamics, Bayesian inference through Markov chain Monte Carlo (MCMC) is a widely used framework owing its popularity to a wide range of available models and its accommodation of phylogenetic uncertainty in generating posterior distributions for all parameters (including the tree topology). In practice however, data set sizes limit the application of Bayesian phylodynamic inference much more than the approaches highlighted earlier in this chapter. Recent applications have involved over a thousand genomes (e.g. for Ebola virus, Dudas et al. (2017)), and one of the largest studies included about 4,000 influenza gene sequences (Bedford et al., 2015). For the latter, strong temporal structure and phylogenetic resolution aided integrating over all plausible evolutionary histories. Many different approaches to confront the computational limitations are currently in development, focusing on different aspects of Bayesian inference.

Typically, MCMC algorithms may suffer from two problems that hamper performance: slow convergence and poor mixing (Nascimento et al., 2017). To remedy this, Bayesian inference often tries to employ a(n approximate) maximum-likelihood tree – which can be quickly estimated (see previous sections) – to yield a better-than-random starting location in tree space to initiate its search. While this aids the search in discrete tree space, it is important to note that other parameters involved in the phylodynamic model are not subject to such a pre-optimisation step and convergence may still take non-negligible time. The continuous need to further optimize MCMC integration is the focus of many current developments in the field.

A related interesting avenue of research is trying to deal with the problem of continuously accumulating data during an ongoing epidemic. The generation of additional sequence data requires an update of previously obtained results, which is typically done with a complete re-evaluation of the integrated Bayesian inference estimation procedure. Such a procedure renders Bayesian approaches costly to maintain an up-to-date estimate of the phylogenetic posterior distribution and this has motivated initial work on "online" Bayesian phylogenetic inference methodology, which can update an existing posterior with new sequences (Dinh et al., 2018; Gill et al., 2020).

While efficiently achieving convergence is one important aspect of the Bayesian challenge, mixing efficiency – which refers to how efficiently the chain samples from the posterior after it has converged on the posterior distribution (Nascimento et al., 2017) – is also of critical importance. In practice, mixing efficiency is frequently assessed by determining the degree of autocorrelation in the MCMC sample, with high autocorrelation reflecting a poor sample to characterise the posterior. If the Markov chain can be made more efficient in

sampling from the posterior, a relatively shorter chain may provide an acceptable estimate of the parameters of interest. Both the model parameterization (and prior specification) and the transition kernels acting upon the model's parameters can have a great effect on mixing efficiency. To deal with the issue of mixing among the potentially vast collection of parameters stemming from different models, adaptive MCMC approaches can potentially increase sampling efficiency for many continuous parameters simultaneously (see Section 3.2).

While convergence and mixing issues will impact the number of MCMC iterations needed to appropriately sample from the posterior, overall computation time will also be determined by the time it takes to evaluate the joint posterior density at each step. The same densities need to be estimated repeatedly and there are no shortcuts to evaluate each observed data likelihood and prior density, no matter which software framework or computational library is being employed. However, developments in multi-core computational hardware – both in the area of traditional central processing units (CPUs) but also of graphical processing units (GPUs) – offer increasing capabilities to compute those densities more efficiently by leveraging high-performance computational libraries. Such libraries provide access for multiple inference software packages to the underlying state-of-the-art hardware, allowing each inference program to avoid implementing low-level access to such hardware. In Chapter 5.4 (Ayres et al. 2020), we describe the BEAGLE high-performance computational library – which is used by multiple phylogenetic inference software packages – to illustrate how such performance increases in computing observed data likelihoods are brought about. All of the approaches described here that deal with Bayesian phylodynamic inference are available through BEAST v1.10 (Suchard et al., 2018), which now by default requires the BEAGLE library to run (Ayres et al., 2019).

3.1 Bayesian phylodynamic inference using BEAST 1.10

While many software packages are available today that focus on phylogenetic and even phylogenomic inference, BEAST (Bayesian Evolutionary Analysis by Sampling Trees; Suchard et al. 2018) unifies molecular phylogenetic reconstruction with complex discrete and continuous trait evolution and allows modelling parameters of interest as a function of external covariate data. In particular, the use of location data associated with genetic sequences has been popularised through Bayesian inference approaches for ancestral location reconstruction, allowing to track the spread of an organism through geographic space.

Since its inception, BEAST has focused on estimating time-scaled (rooted) trees from genetic data. This can be done through the classical use of external calibration information, such as information from the fossil record or through studies on plate tectonics as well as the use of the sampling times of sequences from 'measurably evolving populations' (MEPs, Drummond et al. 2003b). MEPs are characterized by either fast substitution rates and sample availability over a limited time-scale (e.g. for rapidly evolving RNA viruses) or slower substitution rates but with much longer sampling time scales (e.g. ancient DNA, Drummond et al. 2003b). Prior distributions over time-measured genealogies such as the coalescent allow inferring temporal changes in population size. To this end, BEAST provides a wide range of molecular clock models and coalescent models, alongside the traditional range of nucleotide, codon and amino acid substitution models.

In recent years, BEAST has increasingly focused on the analysis of rapidly evolving pathogens and their evolutionary and epidemiological dynamics. Given the rapid growth of pathogen genome sequencing as part of public health responses to infectious diseases, recent areas of interest for further development of BEAST are the exploitation of increasingly parallel computing architectures to decrease time to results, such as multi-core CPU and GPU

hardware, both through the development of novel estimation procedures (such as adaptive MCMC; see Section 3.2) and a much closer integration with the BEAGLE high-performance computational library (see Chapter 5.4 [Ayres et al. 2020] for more information).

3.2 Adaptive MCMC

Novel sequencing technologies are delivering increasingly larger numbers of genome sequences, which may amount to thousands of sequences containing hundreds or even thousands of genes. Viruses with limited genome sizes are typically characterised by a restricted number of genes. In a viral outbreak setting, recent years have witnessed the deployment of portably sequencing technologies (Quick et al., 2016), for example to analyse a large-scale Ebola virus outbreak in West Africa (Dudas et al., 2017) and the Zika epidemic in the Americas (Faria et al., 2017). The increasing reliability and accuracy of portable genome sequencing technologies have now turned them into an important instrument in shedding light on unfolding epidemics.

Large viral data sets can typically be analysed using evolutionary models that take into account the structural properties of those alignments by employing gene-specific and/or codon position-specific partitioning schemes. Such modelling approaches combine the benefits of more accurately modelling the underlying evolutionary processes with increased computational performance, for example by using different nucleotide substitution models per codon position rather than computationally demanding full codon models (Shapiro et al., 2006). However, as partitioning strategies involve estimating conditionally independent models of molecular evolution for different genes and different positions within those genes, they require a large number of evolutionary parameters to be estimated, which may pose difficulties for traditional Bayesian inference approaches. Given the predominance of single-component Metropolis-Hastings approaches, a parameter estimation strategy which proposes a new value for one single parameter at a time (Gilks et al., 1996), estimating large numbers of parameters that are spread across multiple data partitions is associated with a considerable computational burden in Bayesian phylogenetic inference (see Figure 8).

Baele et al. (2017a) have introduced a transition kernel into BEAST (Suchard et al., 2018) based on the adaptive Metropolis (AM) algorithm of Haario et al. (2001) and Roberts and Rosenthal (2009) that continuously adapts its d-dimensional proposal distribution to better match the target distribution and hence learn better parameter values as the analysis progresses. Computing the covariance of the proposal distribution using all of the previous states, the AM algorithm is in turn based on the classical random walk Metropolis algorithm (Metropolis et al., 1953) and earlier work by Haario et al. (1999) that entertains a Gaussian proposal distribution centered on the chain's current state but with the covariance being calculated from a fixed number of previous states. The use of simple recursion formulae to update the covariances ensures that the computational cost associated with the AM algorithm remains constant as the analysis progresses (Haario et al., 2001).

Apart from updating the proposal distribution by using currently available knowledge about the target distribution, the construction of the AM algorithm is identical to the usual random walk Metropolis-based chain. Suppose that at iteration n we have previously sampled the states $X_0, X_1, \ldots, X_{n-1}$, where X_0 is the initial state (typically sampled at random from its prior distribution). A candidate point Y is then sampled from the (asymptotically symmetric) normal proposal distribution, given at iteration n by $Q_n(x, \cdot) = N(x, (C_d)^2 I_d/d)$ for $n \leq C_0$, while for $n > C_0$:

$$Q_n(x, \cdot) = (1 - \beta)N(x, \Sigma_n/d) + \beta N(x, (C_d)^2 I_d/d), \tag{3}$$

where Σ_n is the current empirical estimate of the covariance structure of the target distribution

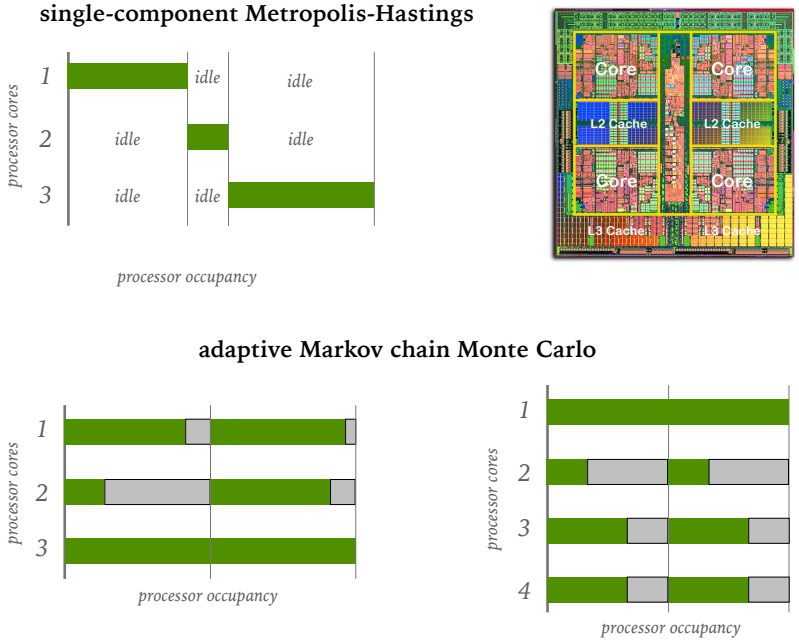


Figure 8 Conceptual visualisation of the potential benefits of an adaptive MCMC algorithm over single-component Metropolis-Hastings when assuming a codon partitioned substitution model (green bars indicate a processor's occupancy while computing a specific likelihood). In Bayesian phylogenetics, the common practice of updating a single parameter at a time typically only requires a single CPU core in order to recompute the observed data likelihood, leaving many CPU cores idle and hence underusing the computational capacity of multi-core CPU architectures. In many cases, the likelihood for the second codon position is quickly computed given the low number of unique site patterns at this position, whereas the third codon position typically accumulates more substitutions resulting in a more demanding likelihood computation. Adaptive MCMC allows updating a collection of continuous parameters simultaneously, putting many cores to work in a parallel fashion to compute the various (codon) partitions. Note that different computational demands between data partitions will lead to waiting times (shown in grey) which hamper performance, a problem that can be tackled by splitting the computation of a particular partition over multiple processor cores. In this example, a fourth processor core is not being used and hence partially offloading the computation of the third codon position likelihood to the remaining free processor core reduces waiting time and increases overall throughput. Quad-Core AMD Opteron processor silicon die shown courtesy of Advanced Micro Devices, Inc. (AMD), obtained from Wikimedia Commons.

based on the run so far, I_d is the d-dimensional identity matrix and β is a small positive constant. The (order of) magnitude for the parameters C_0 and C_d is not easily determined however and, given the only recent introduction of such adaptive transition kernels in Bayesian phylogenetics, is subject to further research. The candidate point Y proposed by the transition kernel is accepted with probability

$$\alpha(X_{n-1}, Y) = \min\left(1, \frac{\pi(Y)}{\pi(X_{n-1})}\right), \tag{4}$$

in which case we set $X_n = Y$, and otherwise $X_n = X_{n-1}$, where $\pi(\cdot)$ is the target distribution.

Note that the chosen probability for the acceptance resembles the familiar acceptance probability of the Metropolis algorithm, but the corresponding stochastic chain is no longer Markovian (Haario et al., 2001). It is known that adaptive MCMC algorithms will not always preserve stationarity of $\pi(\cdot)$ (Roberts and Rosenthal, 2009). However, having proven the ergodicity of adaptive MCMC under certain conditions (Roberts and Rosenthal, 2007), the AM algorithm above will indeed converge to $\pi(\cdot)$ and satisfy the Weak Law of Large Numbers (WLLN), even though it is not Markovian.

The use of a d-dimensional proposal distribution through such an adaptive transition kernel can correspond in its simplest (or standard) use case to updating one parameter in each of d sequence data partitions (e.g. when d genes are present in the multiple sequence alignment). The simultaneous proposal of updated parameter values for all d parameters will in such a case trigger d likelihood calculations, which can be performed in parallel as there is no dependency of the sequence data likelihoods on one another. The current surge in development and availability of multi-core processors, both in the central processing unit (CPU) and graphics processing unit (GPU) processor markets, provides an excellent opportunity to perform such massively parallel computations. Multi-core CPU systems allow evaluating multiple data likelihoods simultaneously on a single processor, employing each processor core to compute the likelihood of a given data partition. GPU cards aimed at the scientific computing market benefit from a different approach towards likelihood computation, with an implementation that is agnostic of the concept of tree topologies (Suchard and Rambaut, 2009). High-performance computational libraries such as BEAGLE (Ayres et al., 2019) provide an extended API and library to support concurrent computation, not only on CPU but also on GPU, of independent partial likelihoods, for increased performance of analyses with greater flexibility of data partitioning (see the separate BEAGLE chapter for more information).

3.3 Discrete phylogeographic inference

Apart from the availability of many nucleotide data partitions in modern-day data sets that need to be analysed using phylogenetic approaches, an increasing number of trait data partitions are being included into phylodynamic analyses (for an overview, see Baele et al. 2017b). Given the specific properties of the data set we analyse in this chapter, we focus here on a discussion of discrete trait analysis in combination with sequence data. As in the previous sections in this chapter, we consider sampling location as a discrete trait to perform phylogeographic analyses in conjunction with approaches to visualize the reconstructed viral spread over time and space.

Discrete phylogeographic inference has witnessed a surge in popularity after its development and inclusion into the BEAST software package (Lemey et al., 2009; Suchard et al., 2018). A popular approach for discrete trait modeling is to take guidance from standard phylogenetics and borrow the process of exchange between sequence character states as a

generic model for how (discrete) traits evolve over the branches of a phylogeny, as described in the previous section.

With most general models, it requires estimating large migration rate matrices and therefore significant computer power and typically makes the resulting analysis no longer suitable even for multi-core CPU systems and requires the use of state-of-the-art GPUs (Suchard and Rambaut, 2009).

Informing the instantaneous rate parameters of a discrete trait model can be aided by providing predictors or covariates that inform the transition rates between discrete states (see Section 3.4). Other approaches to protect against over-parameterization include prior specification, which can for example consist of proposing higher rates of diffusion between nearby locations a priori. In addition, Bayesian stochastic search for variable selection (BSSVS) can be adopted to reduce the number of rate parameters to a restricted set that provides the most adequate parsimonious description of the diffusion process (Lemey et al., 2009). Variable selection and informative prior specification both increase statistical efficiency, which becomes even more important when drawing inference from sparse data – such as a single column of discrete traits – under more complex models, for example, assuming asymmetric transition rates between each pair of discretized trait values (Edwards et al., 2011).

3.4 Incorporating potential predictors of spatial spread

As mentioned in the previous section, the estimation of a potentially large number of transition rates in a discrete trait model can be informed by incorporating predictors that may play an important role in the underlying transition process between trait states. Such predictors can be integrated using a generalized linear model (GLM) formulation for the transition rates, a common approach in statistics that allows to model the linear relationship between a dependent variable (in this case the transition rates) and a collection of independent variables. To identify the relevant subset of predictors out of a number of explanatory variables, the GLM model can be extended with a BSSVS procedure. To this end, the $(K-1) \times K$ parameters that model the instantaneous rates of spread between the K discrete locations $\Lambda_{ij}(\forall i \neq j)$ is modelled as a log linear function of the set of P predictors $(x_1, \ldots, x_P)$ so that

$$\log\Lambda_{ij} = \beta_1\delta_1 x_{i,j,1} + \beta_2\delta_2 x_{i,j,2} + \ldots + \beta_P\delta_P x_{i,j,P}, \tag{5}$$

where $(\beta_1, \ldots, \beta_P)$' represent the effective sizes (or coefficients) for the predictors, quantifying their contribution to Λ, and $(\delta_1, \ldots, \delta_P)$ are (0,1)-indicator variables that govern the inclusion or exclusion of the P predictors in the model. The incorporation of indicator variables allows to perform the BSSVS procedure, which involves letting the data decide whether or not the corresponding predictor provides a significant contribution to the model and hence should be kept as part of the model. In other words, when an indicator δ_p equals 1, then predictor x_p is included in the model, and assessing its effect size through β_p allows interpreting the direction and magnitude of that predictor's contribution. The average value of each indicator across iterations provides an estimate of the inclusion probability of a predictor and can be used to compute a Bayes factor expressing how much the data change our prior opinion about the inclusion of each predictor. Completing the model's specification is done by assuming a small prior probability on each predictor's inclusion that reflects a 50% prior probability on no predictors being included, but specifying equal prior probability on each predictor's inclusion and exclusion yields highly similar results. The Bayes factor

for a predictor BF_p is then calculated by dividing the posterior odds for the inclusion of a predictor with the corresponding prior odds

$$\mathrm{BF}_p = \frac{\mathrm{pp}_p}{1-\mathrm{pp}_p} / \frac{\mathrm{qp}_p}{1-\mathrm{qp}_p}, \tag{6}$$

where pp_p is the posterior probability that predictor p is included, in this case the posterior expectation of indicator δ_p, and qp_p is the prior probability that $\delta_p = 1$.

3.4.1 Predictor data

In order to offer an explanation for the geographic dispersal of emerging pathogens, different sources of information can be incorporated in the phylogeographic testing procedures. To illustrate the principle behind this approach, we have collected two data matrices that we use here as predictors for the geographic spread patterns of dengue. First, we consider air transportation data between the 64 countries from which the dengue sequences were obtained, thereby aggregating data from a passenger flux matrix that quantifies the number of passengers traveling between each pair of airports on a daily basis. We use a dataset provided by OAG (Official Airline Guide) Ltd. (`http://www.oag.com`), containing 4.092 airports and the number of seats on scheduled commercial flights between pairs of airports during the years 2004-2006. We take the number of seats on scheduled commercial flights from airport i to j to be proportional to the number of passengers traveling. Because passenger flux does not differ in a statistically significant manner from symmetry in the global air transportation network (Woolley-Meza et al., 2011), we consider flows that were symmetrized. These air transportation data were converted to "effective distances", which aim to reflect the idea that a small fraction of traffic is effectively equivalent to a large distance, and vice versa (Brockmann and Helbing, 2013). A second predictor consists of the average distances between the locations in our data set. Specifically, we considered the average great-circle distance between two locations based on the pairwise distances between all pairs of airports from the two locations. While further predictors can be added into the GLM, we use these two predictors described here to showcase their use and interpretation.

3.4.2 Results

In our BEAST analysis – comprising 200 million iterations – both predictors were consistently included in the GLM model (the associated indicator variables remained 1, so $E[\delta] = 1$), indicating that our data support an important contribution of these two predictors to the geographic spread of dengue. We also find effect sizes that are centered around negative values, indicating an inverse relationship between the predictors and the instantaneous rates of dengue spread between locations, and credible intervals that exclude 0 ($\beta = -0.71[-0.83, -0.59]$ for air flux and $\beta = -0.49[-0.60, -0.37]$ for geographic distance). In other words, the closer two locations are to one another and the smaller their 'effective distances' are (or the more frequent air travel between them), the more intense the spread of dengue between those two locations becomes. In summary, using a framework that estimates the migration history of dengue while simultaneously testing and quantifying potential predictive variables of spatial spread, we show that the global dynamics of dengue are potentially driven by a combination of air passenger flow and geographic distance between the locations from which these dengue samples originated. However, we caution against drawing strong conclusions from this analysis as only two predictors were tested and considerable bias may exist in

the sampling by country. We note that in addition to offering a phylogeographic testing approach, the GLM parameterisation of discrete trait diffusion also considerably reduces the number of parameters to be estimated. While the standard CTMC model has transition rate parameters that scale quadratically with the number of states, the GLM-diffusion parameters scale linearly with the number of predictors. Although the parameters remain restricted in this way, the likelihood calculation for high state spaces remains associated with a large computational burden. However, this can to a large extent be mitigated by parallelisation as discussed in the BEAGLE chapter.

3.5 Tree visualisation using FigTree

We first focus on presenting the inferred time-stamped phylogenetic tree as one of the key outcomes of our joint inference of sequence and trait data, which includes the parameterization of the geographic spread between location as a function of our predictor data. To this end, we employ FigTree, a popular cross-platform graphical tree display software package. Although it can be used as a general tree visualisation tool, it is particularly powerful to display annotated trees produced by BEAST. In order to construct a maximum clade credibility (MCC) tree, i.e. the single tree in the posterior sample with the largest sum of posterior probabilities across its constituent bifurcations, we first use TreeAnnotator (which is part of the BEAST software package) to summarize the trees from the posterior distribution collected during an analysis that comprised 200 million iterations (after removing the necessary burn-in). The resulting time-stamped MCC tree can then be visualised in different manners and with different annotations in FigTree, as illustrated in Figure 9, where different colours have been used for the clusters of sequences corresponding to the different dengue subtypes.

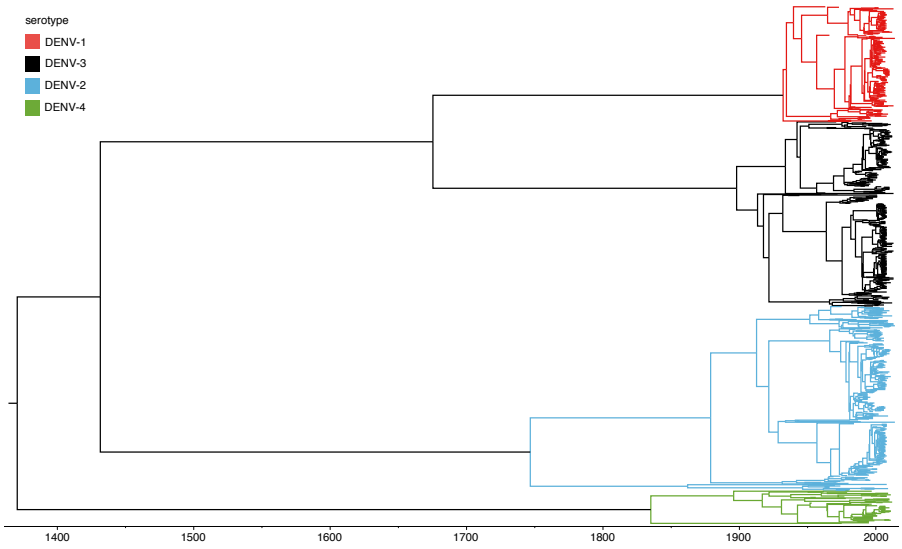


Figure 9 Time-stamped tree visualisation according to the default view in FigTree, with clusters coloured according to the corresponding dengue serotype.

Oftentimes, discrete ancestral trait reconstructions on a tree are represented by branch colour annotations. Such a visualisation allows interpreting the time and location of origin as well as the introduction into different locations at different points in history of dengue. Given that the different dengue serotypes diverged centuries ago (see Figure 9), we here focus on showing the phylogeographic reconstructions of the different serotypes in Figures

10, 11, 12 and 13 (all trees are MCC trees). In doing so, we have "ladderized" the trees, meaning that the nodes have been sorted on one level by the count of their subnodes (on all levels under the node), for easier visualisation and interpretation of the trees.

Dengue virus comprises four serotypes that co-circulate in tropical regions and the relationship between the various dengue serotypes depicted in Figures 9 is well known. It has been hypothesised that antibody-dependent enhancement (ADE) may explain this particular shape of the dengue virus phylogeny, in which the four serotypes are phylogenetically equidistant (see e.g. Grenfell et al. 2004). Natural selection may favour this level of antigenic dissimilarity, as cross-protective antibodies would neutralize more similar strains, whereas more divergent strains would not stimulate ADE. It has also been suggested that independent cross-species (monkey-human) transmission might be able to explain the dengue phylogeny if the serotypes predominate in different geographic areas, followed by later mixing (Holmes and Twiddy, 2003). However, this is difficult to determine based on the visualisation of the trees in Figures 10-13, and a projection onto geographic space would be more useful in such a case, which we discuss in the following section.

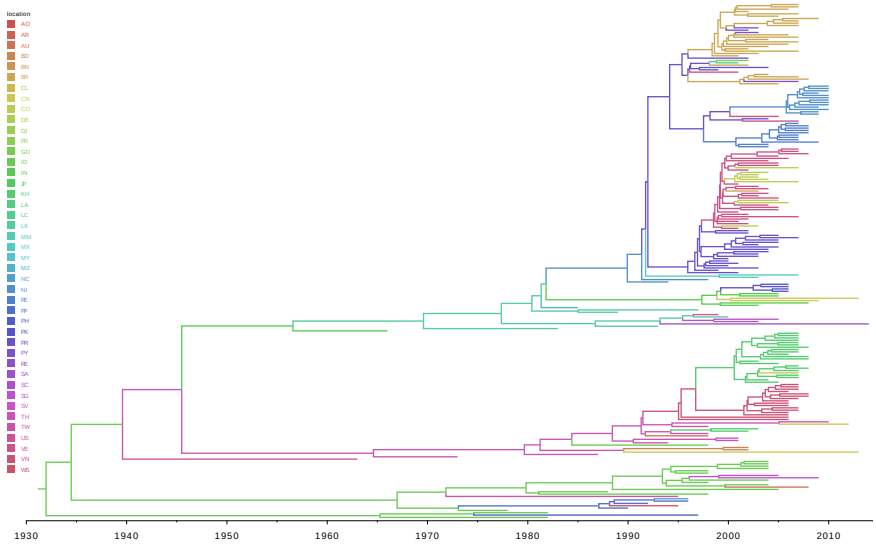


Figure 10 Time-stamped tree visualisation of serotype 1 – based on the full MCC tree (see Figure 9) – with each branch being coloured according to the location state at its descendant node. Based on our Bayesian inference, serotype 1 is estimated to have originated in Indonesia in the 1930s.

3.6 Visualisation and animation using spreaD3

In order to visualise the estimated spread over time in a geographically explicit way, each discrete location considered in the ancestral reconstruction needs to be complemented with a set of GPS coordinates. We here resort to the centroid of each country, which can be easily retrieved online (an alternative option could be to use the GPS coordinates for the capital of the country). These latitude and longitude data can be readily loaded in spreaD3 (Bielejec et al., 2016), along with a geographic map in GeoJSON format containing the regions of interest. SpreaD3 will then generate, by using the sampling times and locations, and the estimated divergence times and ancestral location reconstruction, an animated visualisation of the viral spread over time, starting from the estimated time to most recent common

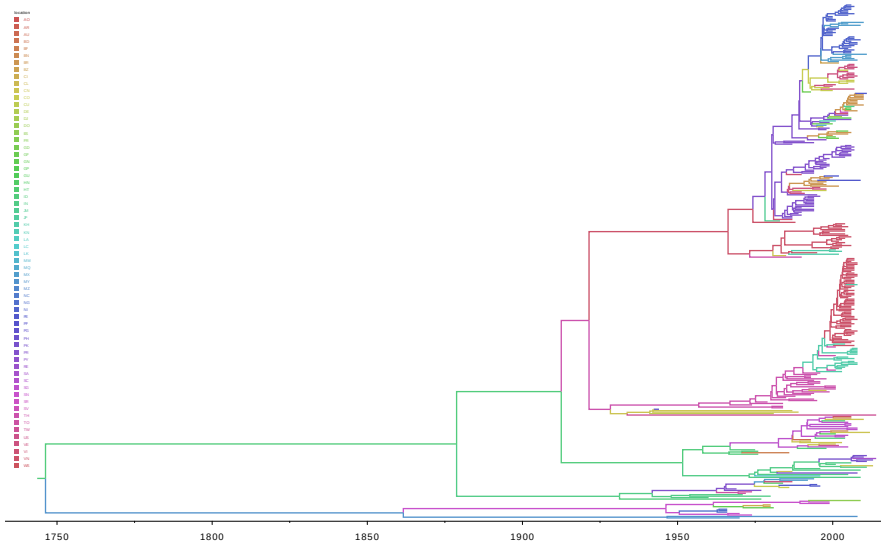


Figure 11 Time-stamped tree visualisation of serotype 2 – based on the full MCC tree (see Figure 9) – with each branch being coloured according to the location state at its descendant node. Based on our Bayesian inference, serotype 2 is estimated to have originated in Indonesia in the first half of the 18th century.

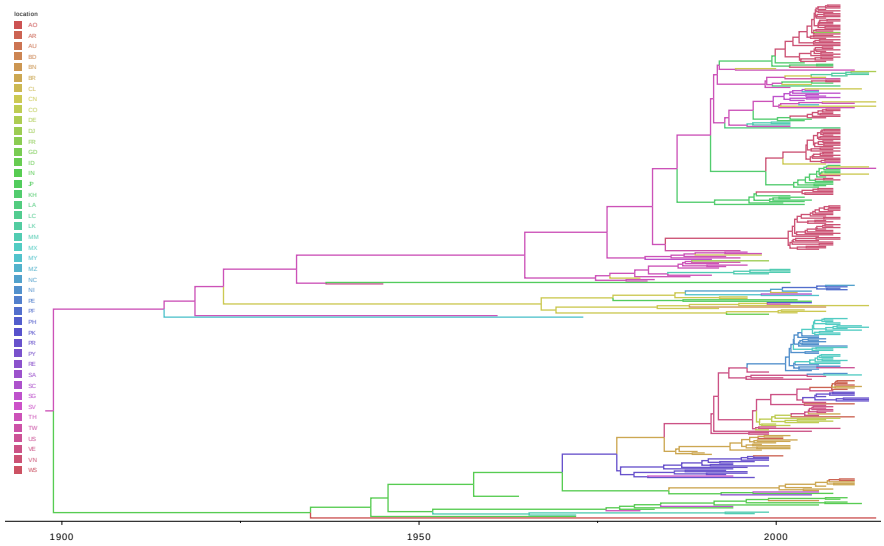


Figure 12 Time-stamped tree visualisation of serotype 3 – based on the full MCC tree (see Figure 9) – with each branch being coloured according to the location state at its descendant node. Based on our Bayesian inference, serotype 3 is estimated to have originated in Thailand at the end of the 19th century.

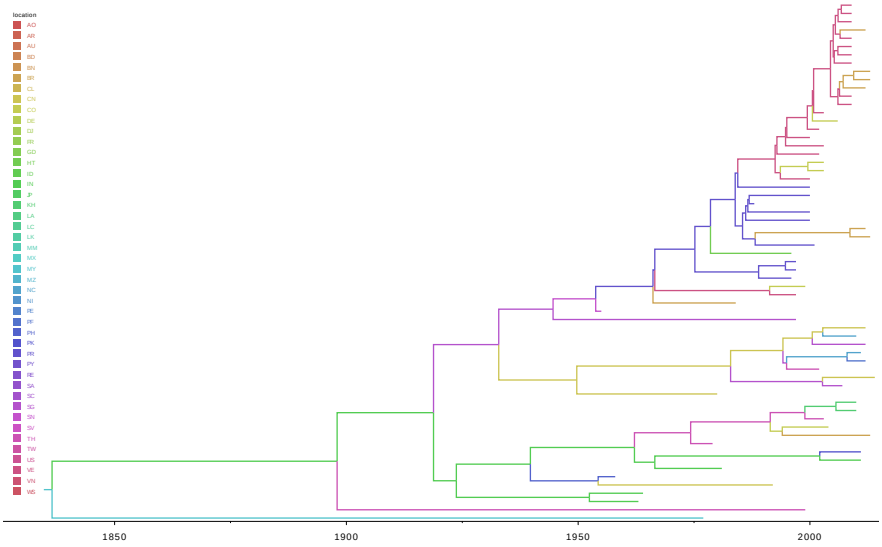


Figure 13 Time-stamped tree visualisation of serotype 4 – based on the full MCC tree (see Figure 9) – with each branch being coloured according to the location state at its descendant node. Based on our Bayesian inference, serotype 4 is estimated to have originated in Myanmar in the first half of the 19th century.

ancestor of the data being analysed, until the most recent sampling date. This is shown in Figures 14-17, where – for each serotype – we present two snapshots of the animation in progress: one taken early on in the epidemic, when dengue still seems to be restricted to South East Asia, and another taken at the start of 2014 (i.e. the most recent sampling date in our dengue data set), where the different dengue serotypes have spread throughout most tropical regions on earth.

4 Comparison between ML and Bayesian estimates

In the previous sections, we have described large whole-genome virus data set analyses using ML and Bayesian approaches, and discussed their inference results on an example of Dengue virus. To assess how their results compare, we looked at the three aspects that correspond to the three ML pipeline steps: (1) the reconstructed tree topologies; (2) the node dates of the time-scaled trees; (3) and the geographic predictions. To make the ML analysis performed on a larger 5 132 complete genome data set comparable with the Bayesian and ML analyses on the small dataset, we pruned the resulting phylogenies to keep only the common tips. We also calculated a consensus topology by collapsing the branches corresponding to internal nodes that were not common to all of the three (pruned) trees (ML (small), ML (large) and Bayesian): each internal node was uniquely identified by the set of tips in its subtree. For the Bayesian case we used the MCC tree.

Tree topologies.

To compare the tree topologies we calculated pairwise normalised quartet distances (Estabrook et al., 1985) (with tqDist [Sand et al. 2014]), where 0.0 indicates identical trees and 1.0 corresponds to trees that have no quartet in common. The topologies obtained using the various inference methods were very similar: the closest ones were reconstructed on the same

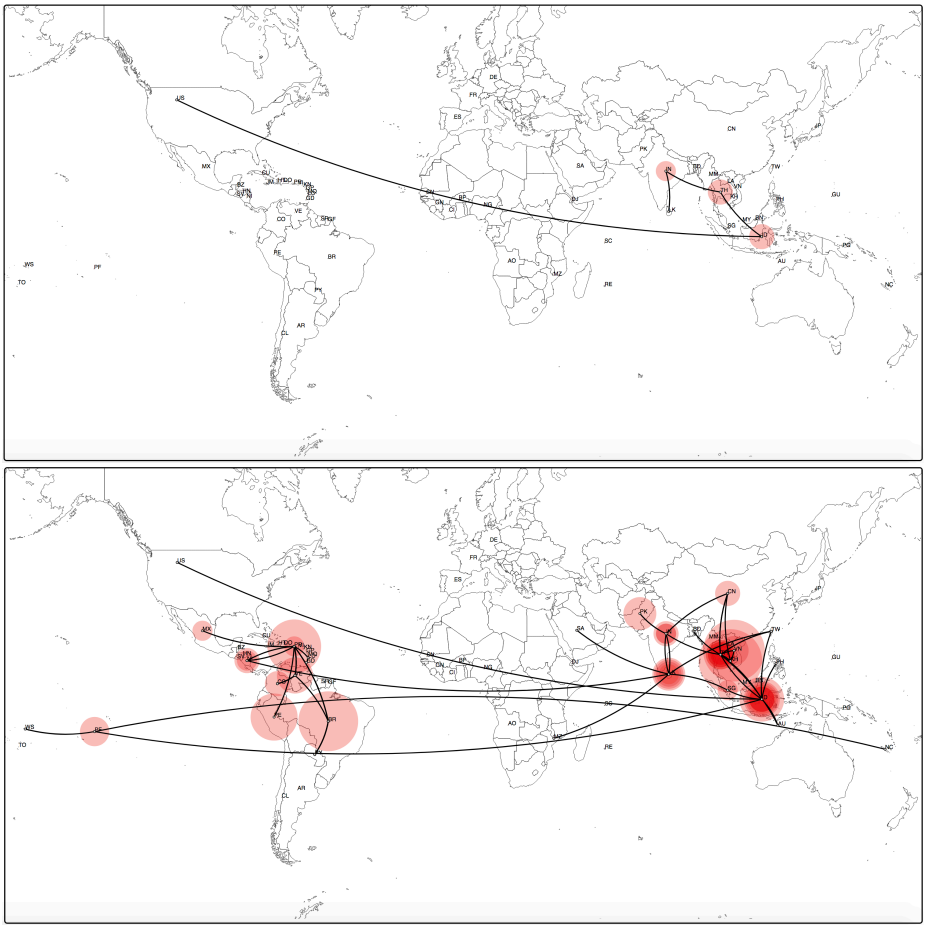


Figure 14 The geographic dispersal over time of dengue serotype 1, visualised using spreaD3 (Bielejec et al., 2016). The top figure shows the spread of dengue serotype 1 at the start of 1970, when the virus is estimated to have originated in Indonesia, from where it first spread to Thailand and the United States. The bottom figure shows the "current" spread (i.e. when the final sample in our data set was taken, at the start of 2014).

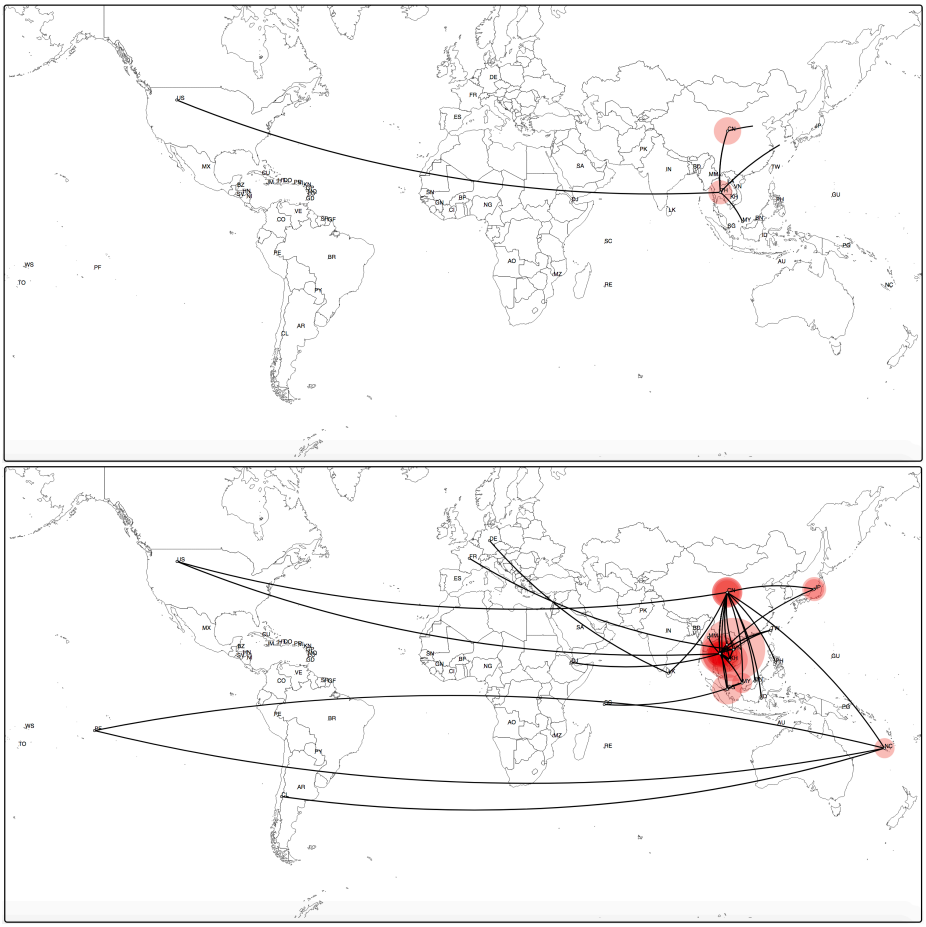


Figure 15 The geographic dispersal over time of dengue serotype 2, visualised using spreaD3 (Bielejec et al., 2016). The top figure shows the spread of dengue serotype 2 at the start of 1980, when the virus is estimated to have originated in Thailand, from where it first spread to China, Myanmar and the United States. The bottom figure shows the "current" spread (i.e. when the final sample in our data set was taken, at the start of 2014).

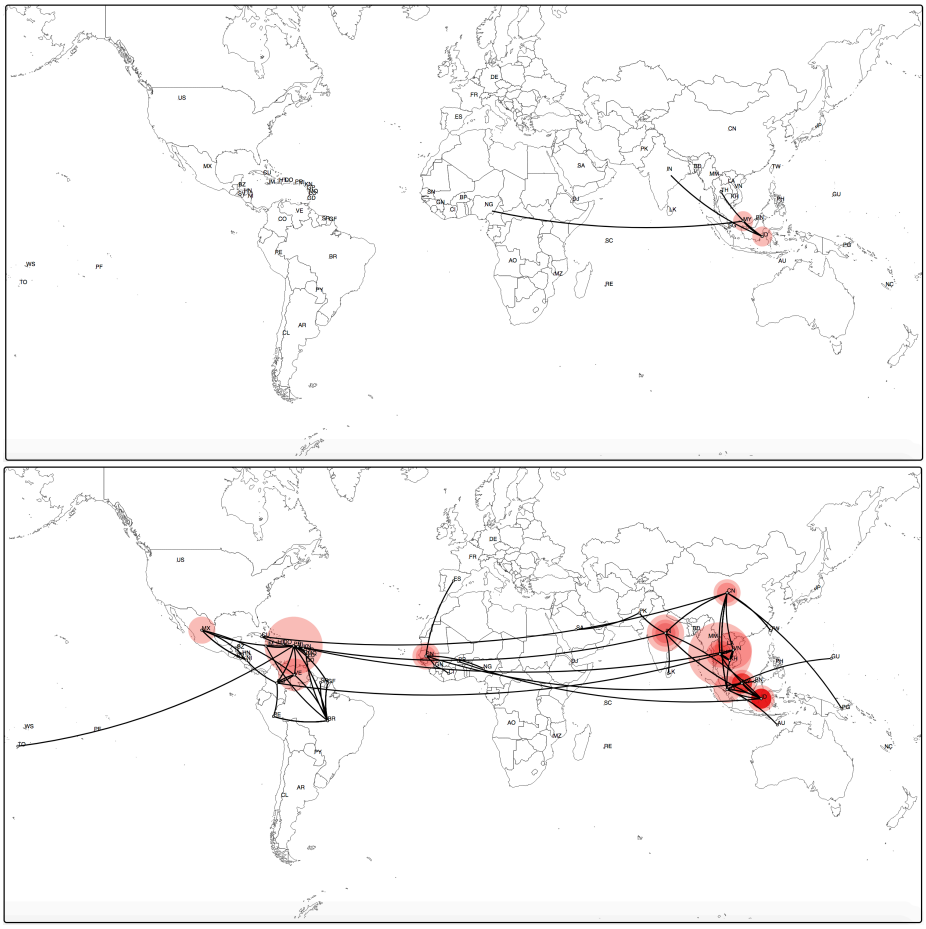


Figure 16 The geographic dispersal over time of dengue serotype 3, visualised using spreaD3 (Bielejec et al., 2016). The top figure shows the spread of dengue serotype 3 at the end of the 1920s, when the virus is estimated to have originated in Indonesia, from where it spread to Thailand, Myanmar, India and the African continent. The bottom figure shows the "current" spread (i.e. when the final sample in our data set was taken, at the start of 2014).

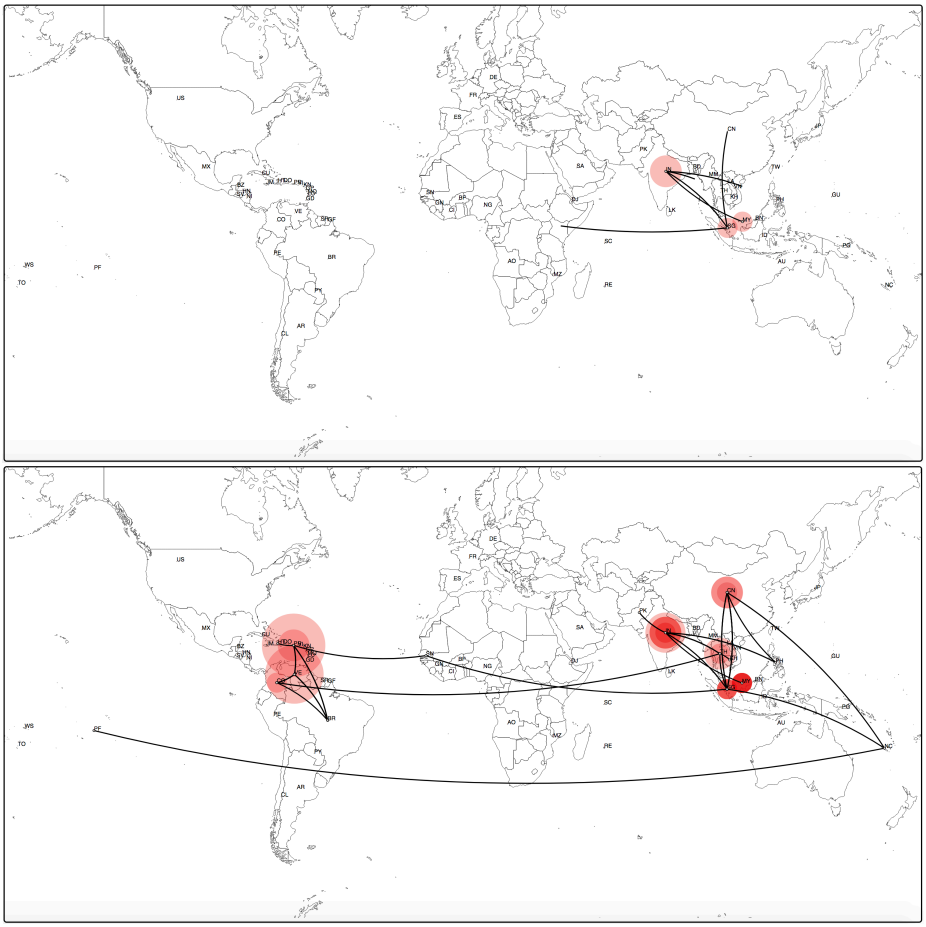


Figure 17 The geographic dispersal over time of dengue serotype 4, visualised using spreaD3 (Bielejec et al., 2016). The top figure shows the spread of dengue serotype 4 at the start of 1950, when the virus is estimated to have originated in Myanmar, from where it spread to India, Myanmar, Singapore and the African continent. The bottom figure shows the "current" spread (i.e. when the final sample in our data set was taken, at the start of 2014).

(small) data set (normalised quartet distance of 0.003), the distance between the two ML trees was 0.009, and 0.011 between the ML tree reconstructed on the large data set and the tree obtained through Bayesian inference.

Time-scaled trees.

We compared the predicted dates of the internal nodes present in the consensus topology. The root (common ancestor of four dengue serotypes) dates were very different and hence strongly depended on the inference used. A root date was estimated to be -2377 $[-2885; -1642]$ for the small ML tree (LSD2) and -54 $[-132; 56]$ for the large one; the Bayesian analysis on the other hand yielded 1370 as the median root date, with a 95% HPD (i.e. $[1215; 1475]$) that can be considered relatively narrow compared to the CIs obtained by various ML tools. These results show that dating relatively deep divergence in a phylogeny can yield drastically different results depending on the methodology used, when all methods rely solely on the sequences and their sampling times, which are all very recent (and given the oldest estimated ages of the full tree may almost be considered to be contemporaneous). This could be predicted from the RTT plot in Figure 4.

We observed considerably smaller root date differences for the serotype subtrees, for which the results are plotted in Figure 18. The time-scaled trees obtained on the same, small, data set are closer to each other than those estimated on different data sets for all the serotypes except for DENV4. In the case of DENV4, the ML time-scaled tree reconstructed on the large data set and the tree from Bayesian inference are very similar. As was already apparent from the full tree root date comparison between the different frameworks, node date estimates obtained through a fully Bayesian inference are consistently more recent than those generated by ML, regardless of the data set size used for the latter. Irrespective of differences in methodology, we stress that estimating relatively deep divergence times based on the divergence accumulating between recent tips can be misleading. Specifically, purifying selection has been shown to lead to severe underestimates of the ancient age of viral lineages (Wertheim and Kosakovsky Pond, 2011). Recent advances in molecular clock modelling attempt to address this (Membrebe et al., 2019).

Geographic reconstruction.

For each internal node in the consensus tree we compared its predicted country and probabilities. As the MPPA method used for ML phylogeographic reconstruction might predict multiple countries per node, we checked whether the modal Bayesian state was among them. For the majority of the nodes it was the case: 98% (91%) for ML ACR on the small (large) data set vs the Bayesian ACR. Less intersection with the large data set could be explained by the fact that it included more countries (83 versus 67): for the ACRs performed on the small data set the 16 unseen countries could never be reconstructed.

5 Conclusions

The combination of high-throughput experimental techniques and advanced methods that stem from physics, statistics and computer science allow to analyse increasingly large quantities of genomic data. ML and Bayesian phylogenomic tools can perform divergence time estimation and phylogeographic reconstruction of thousands of genome-scale virus sequences. However, care should be taken in choosing the data (e.g. removing erroneously annotated and poorly sequenced data that can bias the predictions), choosing correct tools

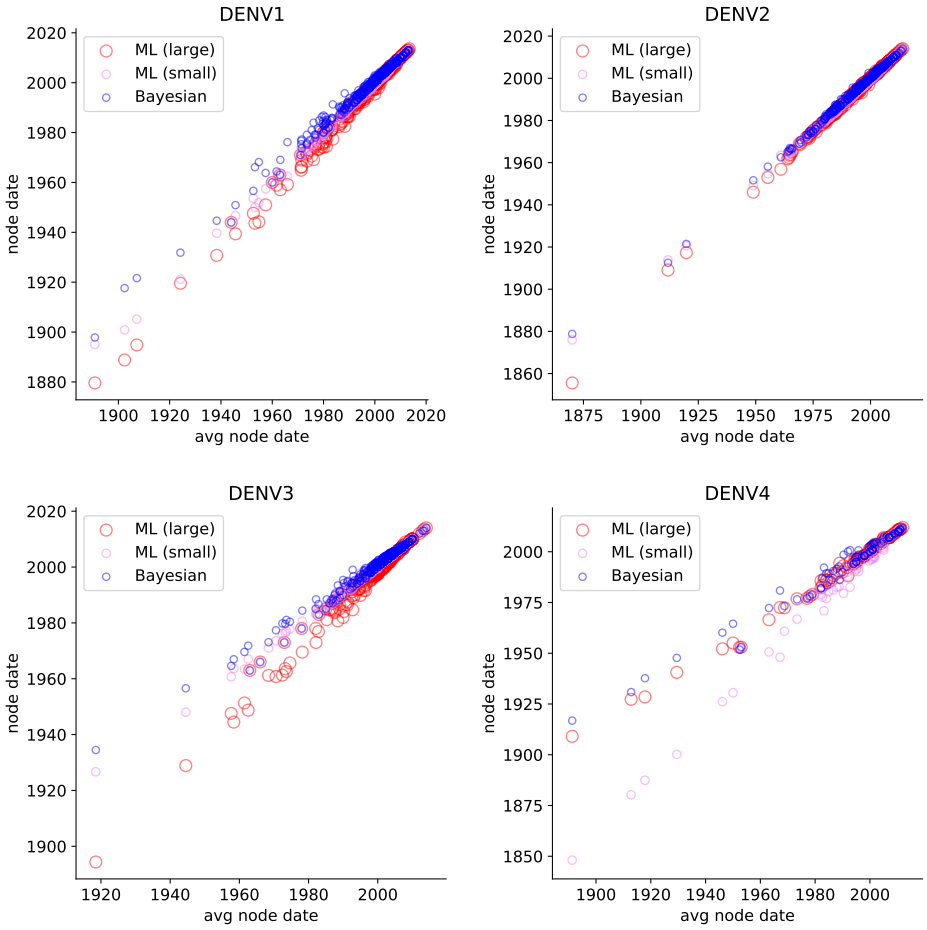


Figure 18 Common node dates for ML (small, dated with LSD2, pink), ML (large, dated with LSD2, red) and Bayesian (blue) trees by serotype (y axes) plotted against the node date averaged over the three trees (x axis): DENV1 (top left), DENV2 (top right), DENV3 (bottom left), and DENV4 (bottom right). Root nodes correspond to the left-most points.

(e.g. FastTree is a good tool for a quick preliminary phylogeny reconstruction but it is not very accurate) and correct configurations (e.g. priors for Bayesian analysis, relaxed vs strict molecular clock), checking the results at all stages of the analysis (e.g. correct tree root position), and comparing the predictions obtained by different methods.

We compared Bayesian and ML analyses on the example of Dengue virus data, and obtained results that are similar overall, but also showed many differences (especially in terms of time predictions). It is however difficult to assess which result is closer to biological truth. ML analysis is much faster to perform ($\sim$ 2 days on a 12-core machine for the large data set and $\sim 2,5$ hours for the small one) and can be therefore applied to larger data sets. However by performing the analysis step by step, it might accumulate error, so each intermediate result needs to be checked. When possible, it may prove useful to perform different analyses and compare the results. The DENV data sets we have used in our comparisons serve to illustrate the computational approaches. We acknowledge that considerable sampling bias may exist between countries, which complicates drawing reliable conclusions about patterns of spatial spread. Furthermore, we have not performed any recombination analyses. As pointed out in the introduction, such analyses should be part of phylogenomic studies of pathogens that may recombine, which is the case for DENV (Worobey et al., 1999).

We note an interesting convergence in the development of Bayesian and ML analysis frameworks. The ML tools for dating trees have evolved towards the inclusion of (1) relaxed molecular clocks, which have to a large extent been advanced in Bayesian frameworks, and (2) statistical tests to decide between strict and relaxed clock models, which also have received much attention in Bayesian frameworks. On the other hand, we illustrated some of the efforts to reduce the computational burden of Bayesian inference in order to decrease the large gap with ML approaches. However, accommodating phylogenetic uncertainty by averaging over all plausible evolutionary histories will always remain restrictive relative to ML estimation.

References

Allen, B. L. and Steel, M. (2001). Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees. *Annals of Combinatorics*, 5(1):1–15.

Ayres, D. L., Cummings, M. P., Baele, G., Darling, A. E., Lewis, P. O., Swofford, D. L., Huelsenbeck, J. P., Lemey, P., Rambaut, A., and Suchard, M. A. (2019). BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst. Biol.*, 68(6):1052–1061.

Ayres, D. L., Lemey, P., Baele, G., and Suchard, M. A. (2020). Beagle 3 high-performance computational library for phylogenetic inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.4, pages 5.4:1–5.4:9. No commercial publisher | Authors open access book.

Baele, G., Lemey, P., Rambaut, A., and Suchard, M. A. (2017a). Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST. *Bioinformatics*, 33(12):1798–1805.

Baele, G., Suchard, M. A., Rambaut, A., and Lemey, P. (2017b). Emerging concepts of data integration in pathogen phylodynamics. *Syst. Biol.*, 66(1):e47–e65.

Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A., Klimov, A., Lewis, N. S., Li, X., McCauley, J. W., Odagiri, T., Potdar, V., Rambaut, A., Shu, Y., Skepner, E., Smith, D. J., Suchard, M. A., Tashiro, M., Wang, D., Xu, X., Lemey, P., and Russell, C. A. (2015). Global

circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523:217–220.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, 41(D1):D36–D42.

Bielejec, F., Baele, G., Vrancken, B., Suchard, M. A., Rambaut, A., and Lemey, P. (2016). Spread3: interactive visualisation of spatiotemporal history and trait evolutionary processes. *Mol. Biol. Evol.*, 33(8):2167–2169.

Blok, J. (1985). Genetic relationships of the dengue virus serotypes. *J. Gen. Virol.*, 66:1323–1325.

Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342:1337–1342.

Bromham, L. (2020). Substitution rate analysis and molecular evolution. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.4, pages 4.4:1–4.4:21. No commercial publisher | Authors open access book.

Bromham, L., Duchêne, S., Hua, X., Ritchie, A. M., Duchêne, D. A., and Ho, S. Y. W. (2018). Bayesian molecular dating: opening up the black box. *Biological Reviews*, 93(2):1165–1191.

Chernomor, O., Minh, B. Q., Forest, F., Klaere, S., Ingram, T., Henzinger, M., Haeseler, A., and Freckleton, R. (2015). Split diversity in constrained conservation prioritization using integer linear programming. *Methods in Ecology and Evolution*, 6(1):83–91.

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, 65(6):997–1008.

Chor, B. and Tuller, T. (2005). Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics*, 21(Suppl 1):i97–i106.

Collins, T. M., Wimberger, P. H., and Naylor, G. J. P. (1994). Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. *Systematic Biology*, 43(4):482.

Dinh, V., Darling, A. E., and Matsen IV, F. A. (2018). Online Bayesian phylogenetic inference: theoretical foundations via sequential monte carlo. *Syst. Biol.*, 67(3):503–517.

Drummond, A., Forsberg, R., and Rodrigo, A. G. (2001). The Inference of Stepwise Changes in Substitution Rates Using Serial Sequence Samples. *Molecular Biology and Evolution*, 18(7):1365–1371.

Drummond, A., Oliver G., Pybus, and Rambaut, A. (2003a). Inference of Viral Evolutionary Rates from Molecular Sequences. *Advances in Parasitology*, 54:331–358.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A. G. (2003b). Measurably evolving populations. *Trends Ecol. Evol.*, 18(9):481–488.

Duchêne, S., Di Giallonardo, F., and Holmes, E. C. (2016a). Substitution Model Adequacy and Assessing the Reliability of Estimates of Virus Evolutionary Rates and Time Scales. *Molecular Biology and Evolution*, 33(1):255–267.

Duchêne, S., Geoghegan, J. L., Holmes, E. C., and Ho, S. Y. (2016b). Estimating evolutionary rates using time-structured data: a general comparison of phylogenetic methods. *Bioinformatics*, 32(22):btw421.

Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J., Ladner, J. T., Arias, A., Asogun, D., Bielejec, F., Caddy, S. L., Cotten, M., D'Ambrozio, J., Dellicour, S., Caro, A. D., Diclaro II, J. D., Durrafour, S., Elmore, M. J., Fakoli III, L. S., Faye, O., Gilbert, M. L., Gevao, S. M., Gire, S., Gladden-Young, A., Gnirke, A.,

Goba, A., Grant, D. S., Haagmans, B. L., Hiscox, J. A., Jah, U., Kargbo, B., Kugelman, J. R., Liu, D., Lu, J., Malboeuf, C. M., Mate, S., Matthews, D. A., Matranga, C. B., Meredith, L. W., Qu, J., Quick, J., Pas, S. D., Phan, M. V. T., Pollakis, G., Reusken, C. B., Sanchez-Lockhart, M., Schaffner, S. F., Schieffelin, J. S., Sealfon, R. S., Simon-Loriere, E., Smits, S. L., Stoecker, K., Thorne, L., Tobin, E. A., Vandi, M. A., Watson, S. J., West, K., Whitmer, S., Wiley, M. R., Winnicki, S. M., Wohl, S., Wölfel, R., Yozwiak, N. L., Andersen, K. G., Blyden, S. O., Bolay, F., Carroll, M. W., Dahn, B., Diallo, B., Formenty, P., Fraser, C., Gao, G. F., Garry, R. F., Goodfellow, I., Günther, S., Happi, C. T., Holmes, E. C., Keïta, S., Kellam, P., Koopmans, M. P. G., Kuhn, J. H., Loman, N. J., Magassouba, N., Naidoo, D., Nichol, S. T., Nyenswah, T., Palacios, G., Pybus, O. G., Sabeti, P. C., Sall, A., Ströher, U., Wurie, I., Suchard, M. A., Lemey, P., and Rambaut, A. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, 544:309–315.

Edwards, C. J., Suchard, M. A., Lemey, P., Welch, J. J., Barnes, I., Fulton, T. L., Barnett, R., O'Connell, T. C., Coxon, P., Monaghan, N., Valdiosera, C. E., Lorenzen, E. D., Willerslev, E., Baryshnikov, G. F., Rambaut, A., Thomas, M. G., Bradley, D. G., and Shapiro, B. (2011). Ancient hybridization and an Irish origin for the modern polar bear matriline. *Curr. Biol.*, 21(15):1251–1258.

Estabrook, G. F., McMorris, F. R., and Meacham, C. A. (1985). Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. *Systematic Biology*, 34(2):193–200.

Faria, N. R., Kraemer, M. U. G., Hill, S. C., Goes de Jesus, J., Aguiar, R. S., Iani, F. C. M., Xavier, J., Quick, J., du Plessis, L., Dellicour, S., Thézé, J., Carvalho, R. D. O., Baele, G., Wu, C.-H., Silveira, P. P., Arruda, M. B., Pereira, M. A., Pereira, G. C., Lourenço, J., Obolski, U., Abade, L., Vasylyeva, T. I., Giovanetti, M., Yi, D., Weiss, D. J., Wint, G. R. W., Shearer, F. M., Funk, S., Nikolay, B., Fonseca, V., Adelino, T. E. R., Oliveira, M. A. A., Silva, M. V. F., Sacchetto, L., Figueiredo, P. O., Rezende, I. M., Mello, E. M., Said, R. F. C., Santos, D. A., Ferraz, M. L., Brito, M. G., Santana, L. F., Menezes, M. T., Brindeiro, R. M., Tanuri, A., dos Santos, F. C. P., Cunha, M. S., Nogueira, J. S., Rocco, I. M., da Costa, A. C., Komninakis, S. C. V., Azevedo, V., Chieppe, A. O., Araujo, E. S. M., Mendonça, M. C. L., dos Santos, C. C., dos Santos, C. D., Mares-Guia, A. M., Nogueira, R. M. R., Sequeira, P. C., Abreu, R. G., Garcia, M. H. O., Abreu, A. L., Okumoto, O., Kroon, E. G., de Albuquerque, C. F. C., Lewandowski, K., Pullan, S. T., Carroll, M., de Oliveira, T., Sabino, E. C., Souza, R. P., Suchard, M. A., Lemey, P., Trindade, G. S., Drumond, B. P., Filippis, A. M. B., Loman, N. J., Cauchemez, S., Alcantara, L. C. J., and Pybus, O. G. (2018). Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*, 361(6405):894–899.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O. G., and Lemey, P. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61.

Faria, R., Quick, J., Morales, I., Thézé, J., Jesus, J., Giovanetti, M., Kraemer, M. U. G., Hill, S. C., Black, A., da Costa, A. C., Franco, L. C., Silva, S. P., Wu, C.-H., Raghwani, J., Cauchemez, S., du Plessis, L., Verotti, M. P., de Oliveira, W. K., Carmo, E. H., Coelho, G. E., Santelli, A. C. F. S., Vinhal, L. C., Henriques, C. M., Simpson, J. T., Loose, M., Andersen, K. G., Grubaugh, N. D., Somasekar, S., Chiu, C. Y., Muñoz-Medina, J. E., Gonzalez-Bonilla, C. R., Arias, C. F., Lewis-Ximenez, L. L., Baylis, S., Chieppe, A. O., Aguiar, S. F., Fernandes, C. A., Lemos, P. S., Nascimento, B. L. S., Monteiro, H. A. O., Siqueira, I. C., de Queiroz, M. G., de Souza, T. R., Bezerra, J. F., Lemos, M. R., Pereira,

G. F., Loudal, D., Moura, L. C., Dhalia, R., França, R. F., Magalhães, T., Marques, E. T., Jaenisch, T., Wallau, G. L., de Lima, M. C., Nascimento, V., de Cerqueira, E. M., de Lima, M. M., Mascarenhas, D. L., Moura Neto, J. P., Levin, A. S., Tozetto-Mendoza, T. R., Fonseca, S. N., Mendes-Correa, M. C., Milagres, F., Segurado, A., Holmes, E. C., Rambaut, A., Bedford, T., Nunes, M. R. T., Sabino, E. C., Alcantara, L. C. J., Loman, N., and Pybus, O. G. (2017). Establishment and cryptic transmission of zika virus in brazil and the americas. *Nature*, 546(7658):406–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., Alpuche-Aranda, C. M., Chapela, I. B., Zavala, E. P., Guevara, D. M. E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C., and The WHO Rapid Pandemic Assessment Collaboration (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. *Science*, 324(5934):1557–1561.

Gascuel, O. and Steel, M. (2014). Predicting the Ancestral Character Changes in a Tree is Typically Easier than Predicting the Root State. *Systematic Biology*, 63(3):421–435.

Gascuel, O. and Steel, M. (2019). A Darwinian Uncertainty Principle. *Systematic Biology*.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Gill, M. S., Lemey, P., Bennett, S. N., Biek, R., and Suchard, M. A. (2016). Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. *Syst. Biol.*, 65(5):1041–1056.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.*, 30(3):713–724.

Gill, M. S., Lemey, P., Suchard, M. A., Rambaut, A., and Baele, G. (2020). Online Bayesian Phylodynamic Inference in BEAST with Application to Epidemic Reconstruction. *Molecular Biology and Evolution*. msaa047.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303:327–332.

Guindon, S., Dufayard, J.-F. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3).

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Statist.*, 14:375–395.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernouilli*, 7(2):223–242.

Hanson-Smith, V., Kolaczkowski, B., and Thornton, J. W. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular biology and evolution*, 27(9):1988–99.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174.

Holmes, E. and Twiddy, S. S. (2003). The origin, emergence and evolutionary genetics of dengue virus. *Infect. Genet. Evol.*, 3(1):19–28.

Hordijk, W. and Gascuel, O. (2005). Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics*, 21(24):4338–4347.

Hué, S., Pillay, D., Clewley, J. P., and Pybus, O. G. (2005). Genetic analysis reveals the complex structure of hiv-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A*, 102(12):4425–9.

Ishikawa, S. A., Zhukova, A., Iwasaki, W., and Gascuel, O. (2019). A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution*, 36(9):2069–2085.

Jiménez-Silva, C. L., Carreño, M. F., Ortiz-Baez, A. S., Rey, L. A., Villabona-Arenas, C. J., and Ocazionez, R. E. (2018). Evolutionary history and spatio-temporal dynamics of dengue virus serotypes in an endemic region of Colombia. *PLOS ONE*, 13(8):e0203090.

Jones, B. R. and Poon, A. F. Y. (2017). node.dating: dating ancestors in phylogenetic trees in R. *Bioinformatics*, 33(6):932–934.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of Protein Molecules. In *Mammalian Protein Metabolism*, pages 21–132. Elsevier.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Kozlov, A. M. and Stamatakis, A. (2020). Using raxml-ng in practice. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.3, pages 1.3:1–1.3:25. No commercial publisher | Authors open access book.

Langley, C. H. and Fitch, W. M. (1974). An examination of the constancy of the rate of molecular evolution. *Journal of molecular evolution*, 3(3):161–77.

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution*, 34(9):2422–2424.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finding its roots. *PLoS Comp. Biol.*, 5(9):e1000520.

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.*, 27(8):1877–1885.

Letunic, I. and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128.

Membrebe, J. V., Suchard, M. A., Rambaut, A., Baele, G., and Lemey, P. (2019). Bayesian inference of evolutionary histories under time-dependent substitution rates. *Mol Biol Evol*, 36(8):1793–1803.

Messina, J. P., Brady, O. J., Scott, T. W., Zou, C., Pigott, D. M., Duda, K. A., Bhatt, S., Katzelnick, L., Howes, R. E., Battle, K. E., Simmons, C. P., and Hay, S. I. (2014). Global spread of dengue virus types: mapping the 70 year history. *Trends Microbiol.*, 22(3):138–146.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091.

Minin, V. M., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.*, 25(7):1459–1471.

Nascimento, F. F., dos Reis, M., and Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nat. Ecol. Evol.*, 1:1446–1454.

Naveca, F. G., Claro, I., Giovanetti, M., de Jesus, J. G., Xavier, J., Iani, F. C. d. M., do Nascimento, V. A., de Souza, V. C., Silveira, P. P., Lourenço, J., Santillana, M., Kraemer, M. U. G., Quick, J., Hill, S. C., Thézé, J., Carvalho, R. D. d. O., Azevedo, V.,

Salles, F. C. d. S., Nunes, M. R. T., Lemos, P. d. S., Candido, D. d. S., Pereira, G. d. C., Oliveira, M. A. A., Meneses, C. A. R., Maito, R. M., Cunha, C. R. S. B., Campos, D. P. d. S., Castilho, M. d. C., Siqueira, T. C. d. S., Terra, T. M., Albuquerque, C. F. C. d., Cruz, L. N. d., Abreu, A. L. d., Martins, D. V., Simoes, D. S. d. M. V., Aguiar, R. S. d., Luz, S. L. B., Loman, N., Pybus, O. G., Sabino, E. C., Okumoto, O., Alcantara, L. C. J., and Faria, N. R. (2019). Genomic, epidemiological and digital surveillance of chikungunya virus in the brazilian amazon. *PLOS Neglected Tropical Diseases*, 13(3):1–21.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Pollett, S., Melendrez, M., Maljkovic Berry, I., Duchêne, S., Salje, H., Cummings, D., and Jarman, R. (2018). Understanding dengue virus evolution to support epidemic surveillance and counter-measure development. *Infection, Genetics and Evolution*, 62:279–295.

Price, M. N., Dehal, P. S., Arkin, A. P., Rojas, M., and Brodie, E. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490.

Pupko, T., Pe, I., Shamir, R., and Graur, D. (2000). A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*, 17(6):890–896.

Quick, J., Loman, N., Durrafour, S., Simpson, J., Severi, E., Cowley, L., Bore, J., Koundouno, R., Dudas, G., Mikhail, A., Ouedraogo, N., Afrough, B., Bah, A., Baum, J., Becker-Ziaja, B., Boettcher, J., Cabeza-Cabrerizo, M., Camino-Sanchez, A., Carter, L., Doerrbecker, J., Enkirch, T., Garcia-Dorival, I., Hetzelt, N., Hinzmann, J., Holm, T., Kafetzopoulou, L., Koropugui, M., Kosgey, A., Kuisma, E., Logue, C., Mazzarelli, A., Meisel, S., Mertens, M., Michel, J., Ngabo, D., Nitzsche, K., Pallasch, E., Patrono, L., Portmann, J., Repits, J., Rickett, N., Sachse, A., Singethan, K., Vitoriano, I., Yemanaberhan, R., Zekeng, E., Racine, T., Bello, A., Faye, O., Faye, O., Magassouba, N., Williams, C., Amburgey, V., Winona, L., Davis, E., Gerlach, J., Washington, F., Monteil, V., Jourdain, M., Bererd, M., Camara, A., Somlare, H., Camara, A., Gerard, M., Bado, G., Baillet, B., Delaune, D., Nebie, K., Diarra, A., Savane, Y., Pallawo, R., Gutierrez, G., Milhano, N., Roger, I., Williams, C., Yattara, F., Lewandowski, K., Taylor, J., Rachwal, P., Turner, D., Pollakis, G., Hiscox, J., Matthews, D., O'Shea, M., Johnston, A., Wilson, D., Hutley, E., Smit, E., Caro, A. D., Wölfel, R., Stoecker, K., Fleischmann, E., Gabriel, M., Weller, S., Koivogui, L., Diallo, B., Keïta, S., Rambaut, A., Formenty, P., Günther, S., and Carroll, M. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232.

Rambaut, A. (2000). Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16(4):395–399.

Rambaut, A., Lam, T. T., Carvalho, L. M., and Pybus, O. G. (2016a). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.*, 2(1):vew007.

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016b). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution*, 2(1):vew007.

Ratmann, O., Wymant, C., Colijn, C., Danaviah, S., Essex, M., Frost, S., Gall, A., Gaseitsiwe, S., Grabowski, M. K., Gray, R., Guindon, S., von Haeseler, A., Kaleebu, P., Kendall, M., Kozlov, A., Manasa, J., Minh, B. Q., Moyo, S., Novitsky, V., Nsubuga, R., Pillay, S.,

Quinn, T. C., Serwadda, D., Ssemwanga, D., Stamatakis, A., Trifinopoulos, J., Wawer, M., Brown, A. L., de Oliveira, T., Kellam, P., Pillay, D., Fraser, C., and on behalf of the PANGEA-HIV Consortium (2017). Hiv-1 full-genome phylogenetics of generalized epidemics in sub-saharan africa: Impact of missing nucleotide characters in next-generation sequences. *AIDS Research and Human Retroviruses*, 33(11):1083–1098.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive MCMC. *J. Appl. Prob.*, 44:458–475.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *J. Comp. Graph. Stat.*, 18:349–367.

Robinson, D. (1971). Comparison of labeled trees with valency three. *Journal of Combinatorial Theory, Series B*, 11(2):105–119.

Rota-Stabelli, O., Lartillot, N., Philippe, H., and Pisani, D. (2013). Serine Codon-Usage Bias in Deep Phylogenomics: Pancrustacean Relationships as a Case Study. *Systematic Biology*, 62(1):121–133.

Sagulenko, P., Puller, V., and Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4(1).

Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., and Pedersen, C. N. S. (2014). tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*, 30(14):2079–2080.

Sanderson, M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 37(Database issue):D5—-15.

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X. L., and Mullins, J. I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology*, 73(12):10489–502.

Shapiro, B., Rambaut, A., and Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.*, 23(1):7–9.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463.

Suchard, M., Lemey, P., Baele, G., Ayres, D., Drummond, A., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):vey016.

Suchard, M. A. and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics*, 25:1370–1376.

Swofford, D., Olsen, G., Waddell, P., and Hillis, D. (1996). Phylogenetic Inference. In Hillis, D., Moritz, C., and Mable, B., editors, *Molecular Systematics*, pages 407–514. Oxford University Press, Incorporated.

Tan, K.-K., Zulkifle, N.-I., Sulaiman, S., Pang, S.-P., NorAmdan, N., MatRahim, N., Abd-Jamil, J., Shu, M.-H., Mahadi, N. M., and AbuBakar, S. (2018). Emergence of the Asian lineage dengue virus type 3 genotype III in Malaysia. *BMC Evolutionary Biology*, 18(1):58.

Telford, M. J. (2007). Phylogenomics. *Current Biology*, 17(22):R945–R946.

To, T.-H., Jung, M., Lycett, S., and Gascuel, O. (2016). Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic biology*, 65(1):82–97.

Vilsker, M., Moosa, Y., Nooij, S., Fonseca, V., Ghysens, Y., Dumon, K., Pauwels, R., Alcantara, L. C., Vanden Eynden, E., Vandamme, A.-M., Deforche, K., and de Oliveira, T. (2019). Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics*, 35(5):871–873.

Volz, E. M. and Frost, S. D. W. (2017). Scalable relaxed clock phylogenetic dating. *Virus Evolution*, 3(2).

Walimbe, A. M., Lotankar, M., Cecilia, D., and Cherian, S. S. (2014). Global phylogeography of Dengue type 1 and 2 viruses reveals the role of India. *Infection, Genetics and Evolution*, 22:30–39.

Wertheim, J. O. and Kosakovsky Pond, S. L. (2011). Purifying selection can obscure the ancient age of viral lineages. *Mol Biol Evol*, 28(12):3355–65.

Woolley-Meza, O., Thiemann, C., Grady, D., Lee, J., Seebens, H., Blasius, B., and Brockmann, D. (2011). Complexity in human transportation networks: a comparative analysis of worldwide air transportation and global cargo-ship movements. *Eur. Phys. J. B.*, 84:589–600.

Worobey, M., Rambaut, A., and Holmes, E. C. (1999). Widespread intra-serotype recombination in natural populations of dengue virus. *Proc Natl Acad Sci U S A*, 96(13):7352–7.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yebra, G., Hodcroft, E. B., Ragonnet-Cronin, M. L., Pillay, D., Brown, A. J. L., Consortium, P. H., and Project, I. (2016). Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. *Scientific Reports*, 6(39489).

Yoder, A. D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.*, 17(1):1081–1090.

Zhang, J. and Nei, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of Molecular Evolution*, 44(S1):S139–S146.

Zhou, X., Shen, X.-X., Hittinger, C. T., and Rokas, A. (2018). Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution*, 35(2):486–503.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. *Evolving Genes and Proteins*, pages 97–166.

Chapter 5.4 BEAGLE 3 High-performance Computational Library for Phylogenetic Inference

Daniel L. Ayres
Center for Bioinformatics and Computational Biology, University of Maryland, USA
ayres@umd.edu

Philippe Lemey
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven – University of Leuven, Leuven, Belgium

Guy Baele
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven – University of Leuven, Leuven, Belgium

Marc A. Suchard
Department of Biomathematics, Department of Biostatistics, Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

Abstract

Maximum-likelihood and Bayesian inference approaches to statistical phylogenetics require repeatedly computing of the observed sequence data likelihood. As increasingly cheaper sequencing technology provides phylogenomic studies with larger data set sizes, there stands a critical need to efficiently evaluate this likelihood function in order for phylogenetic computations to complete in reasonable time. The adoption of powerful computing architectures, in the form of multi-core central processing units and many-core graphics cards dedicated to scientific computing, offers unprecedented opportunity to perform massively parallel computation in many research fields, including likelihood evaluation in phylogenetics. In this chapter, we provide insight into the inner workings of BEAGLE, a high-performance likelihood-calculation platform for use on multi-core and many-core computer systems (ubiquitous nowadays in standard desktop computers and laptops) and available in several phylogenetic inference applications to improve computational performance.

How to cite: Daniel L. Ayres, Philippe Lemey, Guy Baele, and Marc A. Suchard (2020). BEAGLE 3 High-performance Computational Library for Phylogenetic Inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 5.4, pp. 5.4:1–5.4:9. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 BEAGLE 3 high-performance computational library

BEAGLE defines a uniform application programming interface (API) and includes a collection of efficient implementations for evaluating the phylogenetic likelihood function under a wide range of evolutionary models, on multi-core central processing units (CPUs) and, importantly, many-core graphics processing units (GPUs). The BEAGLE library can be installed as a shared resource, to be used by any software aimed at phylogenetic reconstruction that supports the library. This approach allows developers of phylogenetic software packages to share in optimizations of the core calculations and for any program that uses BEAGLE to benefit from improvements to the library. For researchers, this centralization provides a single installation to take advantage of new hardware and parallelization techniques.

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 5.4; pp. 5.4:1–5.4:9

A book completely handled by researchers.
No publisher has been paid.

BEAGLE is typically associated with BEAST, and in fact, recent versions of BEAST require the presence of the BEAGLE library in order to performe phylogenetic and/or phylodynamic inference. This is also the case for the analyses performed in the chapter on "Efficiently analysing large viral data sets in computational phylogenomics," that makes extensive use of the BEAGLE (Ayres et al., 2019) library. Many of its features also are used within MrBayes (Ronquist et al., 2012) and BEAST 2.5 (Bouckaert et al., 2019). Further, support for maximum-likelihood inference has been available since BEAGLE's inception, with significant speedups observed in packages such as GARLI (Zwickl, 2006) and PhyML (Guindon et al., 2010), and with work currently underway for PAUP* (Swofford, 2003).

BEAGLE version 3 (Ayres et al., 2019), the latest release of the library, enables analyses with data partitions or with few site patterns to benefit from significant performance increases on GPUs. It also adds new CPU-threaded and GPU software implementations, allowing more effective utilization of a wide range of modern parallel processors. BEAGLE 3 is free, open-source software licensed under the Lesser GPL and available for Windows, Mac and Linux from `https://github.com/beagle-dev/beagle-lib/releases`.

1.1 Parallel Computation with BEAGLE

Advances in computer hardware, specifically in parallel architectures, such as multi-core CPUs, CPU intrinsics (e.g., SSE, AVX) and many-core GPUs, have created opportunities to speed up computationally intensive methods. The structure of the likelihood calculation, involving large numbers of positions and multiple states, as well as other characteristics, makes it a very appealing computational fit to these modern parallel processors, especially to GPUs. Recognizing that different independent calculations are possible in computing phylogenetic likelihoods, as well as that different computing hardware architectures have different strengths with respect to parallel computation, BEAGLE implements concurrent computation in a variety of ways. These can be roughly categorized into three broad levels based on granularity: fine-, medium- and course-grained parallelism.

1.1.1 Fine-Grained Parallelism

BEAGLE exploits GPUs via fine-grained parallelization of functions necessary for computing the likelihood on a phylogenetic tree. Phylogenetic inference programs typically explore tree space in a sequential manner (Figure 1, *tree space*) or with a small number of sampling chains, thus offering a low upper limit for coarse-grained parallelization. In contrast, the crucial computation of partial likelihood arrays at each node of a proposed tree presents an excellent opportunity for fine-grained data parallelism, for which GPUs are especially suited. The use of many lightweight execution threads incurs very low overhead on GPUs and the presence of large numbers of positions and multiple states enables efficient parallelism at this level (Figure 1, *partial likelihood*).

Furthermore, BEAGLE uses GPUs to parallelize other functions necessary for computing the overall tree likelihood, thus minimizing data transfers between the CPU and GPU. These additional functions include those necessary for computing branch transition probabilities, for integrating root and edge likelihoods, and for summing site likelihoods.

BEAGLE also provides SSE and OpenCL implementations for exploiting fine-grained parallelism on CPUs that vectorize likelihood calculations across characters and character states. These solutions, however, offer only a modest performance benefit as CPU vectorization intrinsics are of limited width (128 bits are available with SSE and up to 512 bits with AVX vectorization). Additionally, CPU architectures have lower memory bandwidth than

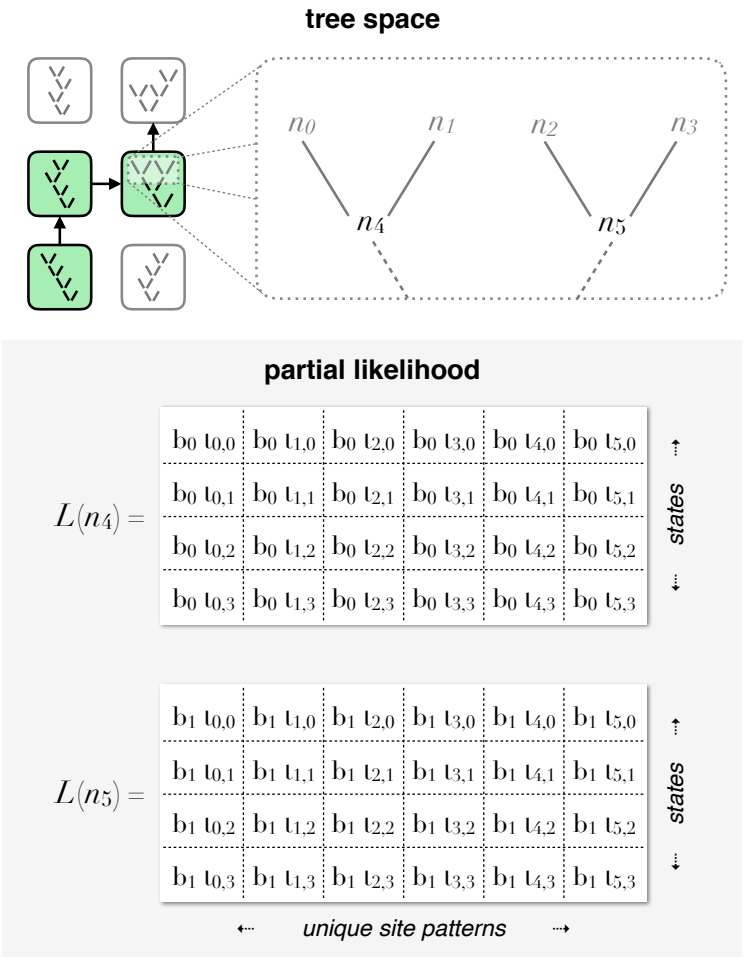


Figure 1 Diagrammatic example of the tree sampling process and medium and fine-grained parallel computation of phylogenetic partial likelihoods using BEAGLE on GPUs for a nucleotide-model problem with 5 taxa, 5 site patterns. Each entry in a partial likelihood array L is assigned to a separate GPU thread t, and each array is assigned to a separate GPU execution block b. In this simplified example, 48 GPU threads are created to enable parallel evaluation of each entry of the partial likelihood arrays $L(n_4)$ and $L(n_5)$.

5.4:4 BEAGLE 3 library

GPUs and we have found this to be a limiting factor when it comes to fine-grained parallel computation of phylogenetic likelihoods.

1.1.2 Medium-Grained Parallelism

In order to calculate the overall likelihood of a proposed tree, phylogenetic inference programs perform a tree traversal, evaluating a partial likelihood array at each node. With BEAGLE, the evaluation of these multi-dimensional arrays is offloaded to the library. Further, when these partial likelihood arrays are independent from one another, they may also be evaluated in parallel to one another, with BEAGLE assigning the calculation of each array to separate execution blocks on the GPU (Figure 1, *partial likelihood*).

With BEAGLE version 3, partitioned analyses also benefit from multi-core CPUs and GPUs by parallelizing the computation of multiple data subsets (Ayres and Cummings, 2017b,a). This capability suits the trend of phylogenomic data sets that are often heavily partitioned in order to better model the underlying evolutionary processes.

1.1.3 Coarse-Grained Parallelism

Phylogenetic inference programs that implement multiple Markov chain Monte Carlo chains or independent runs can invoke multiple BEAGLE library instances, one for each chain or run. For effective parallelism, this is more efficient on multiple hardware resources (e.g., multiple GPUs) and for each library instance to be assigned to a separate resource.

1.2 Library and API Design

1.2.1 Library

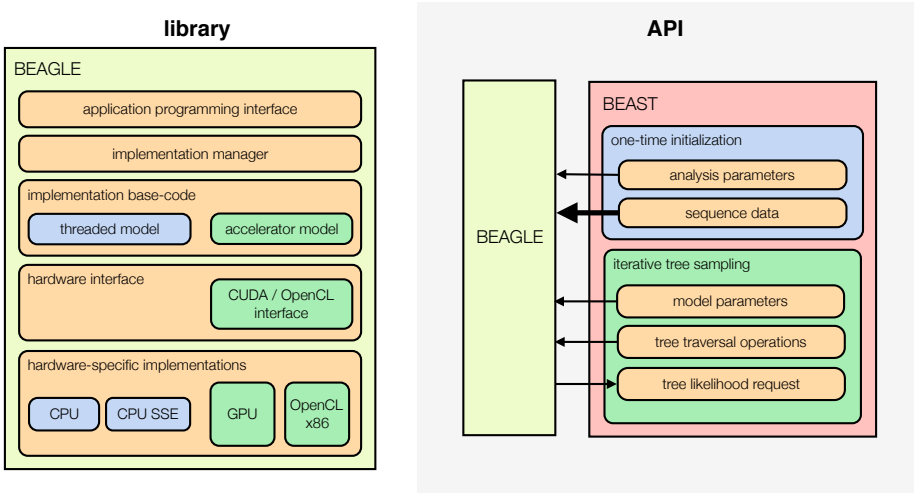


Figure 2 Layer diagrams depicting the BEAGLE library organization, and illustration of API use. Arrows indicate direction and relative size of data transfers between the client program and library.

The general structure of the BEAGLE library can be conceptualized as a set of layers (Figure 2, *library*), the uppermost of which is the application programming interface (API). Underlying this API is an implementation management layer, which loads the available

implementations, makes them available to the client program, and passes API commands to the selected implementation.

The design of BEAGLE allows for new implementations to be developed without the need to alter the core library code or how client programs interface with the library. This architecture also includes a plugin system that allows implementation-specific code (via shared libraries) to be loaded at runtime when the required dependencies are present. Consequently, new frameworks and hardware platforms can more easily be made available to programs that use the library, and ultimately to users performing phylogenetic analyses.

The implementations in BEAGLE version 3 derive from two general models. One is a threaded CPU implementation model that does not directly use external frameworks. Under this model, there is a parallel CPU implementation, and one with added SSE intrinsics that uses vector processing extensions present in many CPUs to further parallelize computation across character state values.

The other implementation model involves an explicit parallel accelerator programming model, and uses the CUDA and the OpenCL external computing frameworks to exploit parallel hardware (Ayres and Cummings, 2017b). It implements fine-grained and medium-grained parallelism for evaluating likelihoods under arbitrary molecular evolutionary models, thus being able to harness large numbers of processing cores to efficiently perform calculations (Suchard and Rambaut, 2009; Ayres et al., 2019).

At the lowest implementation level in BEAGLE, functions that impart a crucial effect on performance are differentiated for each hardware type. This allows for distinctly optimized parallel implementations that are shown in Figure 2, one for NVIDIA and OpenCL-compatible GPUs and one for OpenCL-compatible x86 parallel resources such as multicore CPUs with SIMD-extensions.

1.2.2 Application Programming Interface

The BEAGLE API was designed to increase performance via parallelization while reducing data transfer and memory copy overhead to an external hardware accelerator device (e.g., GPU). Client programs, such as BEAST (Suchard et al., 2018), use the API to offload the evaluation of tree likelihoods to the BEAGLE library (Figure 2, *API*). API functions can be subdivided into two categories: those which are only executed once per inference run and those which are repeatedly called as part of an iterative sampling process. For the one-time initialization process, client programs use the API to indicate analysis parameters such as tree size and sequence length, as well as specifying the type of evolutionary model and hardware resource(s) to be used. This allows BEAGLE to allocate the appropriate number and size of data buffers on device memory. Also at this initialization stage, the sequence data are specified and transferred to device memory. This costly memory operation is only performed once, thus minimizing its impact.

During the iterative tree sampling procedure, client programs use the API to specify changes to the evolutionary model and instruct a series of partial likelihood operations that traverse the proposed tree in order to find its overall likelihood. BEAGLE efficiently computes these operations and makes the overall tree likelihood as well as per-site likelihoods available via another API call.

2 BEAGLE in practice

2.1 Performance

Peak performance with BEAGLE is achieved when using a high-end GPU, with the relative gain over using a CPU depending on model type and problem size as more demanding analyses allow for better utilization of GPU cores. Figure 3 shows speedups relative to single-core CPU code when using BEAGLE on multiple CPU cores and on an NVIDIA Tesla P100 GPU for calculating the likelihood function under a nucleotide model, with increasing unique site pattern counts. Computing the likelihood function typically accounts for over 90% of the total execution time for phylogenetic inference programs and the relationship between speedups and problem size observed here primarily matches what would be observed for a full analysis.

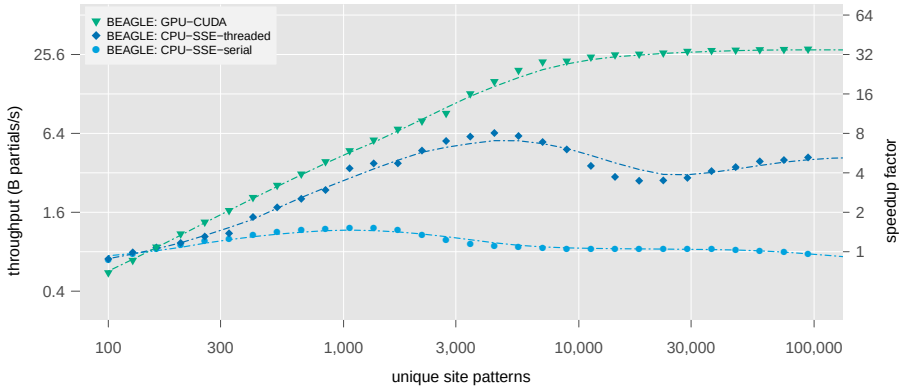


Figure 3 Absolute (throughput in billions of partial likelihood calculations per second) and relative (fold-speedup relative to the slowest performance observed for the BEAGLE library CPU with SSE at any number of unique site patterns) performance for native implementations of the BEAGLE library version 3 on an Intel Xeon E5-2690v4 CPU and NVIDIA Tesla P100 GPU. The data are simulated nucleotide sequences for a tree with 128 tips.

Using a nucleotide model, relative GPU performance over the CPU strongly scales with the number of site patterns. For very small numbers of patterns the GPU exhibits poor performance due to greater execution overhead relative to overall problem size. GPU performance improves quickly as the number of unique site patterns is increased and by $10,000$ patterns it is closer to a saturation point, continuing to increase but with diminishing returns. At $100,000$ nucleotide patterns the GPU is approximately $64\times$ faster than the serial CPU implementation.

For partitions with higher state counts, such as those found in discrete traits models used in phylodynamic analyses, GPU speedup may be achieved with a single character. This is due to the better parallelization opportunity afforded by the increased number of states that can be encoded by the character. The higher state count of such data compared to nucleotide data increases the ratio of computation to data transfer, resulting in increased GPU performance for each character.

2.2 Memory usage

When assessing the suitability of GPU acceleration via BEAGLE for a phylogenetic analysis, it is also important to consider if the GPU has sufficient on-board memory for the analysis to be performed. GPUs typically have less memory than what is available to CPUs and the high transfer cost of moving data from CPU to GPU memory prevents direct use of CPU memory for GPU acceleration.

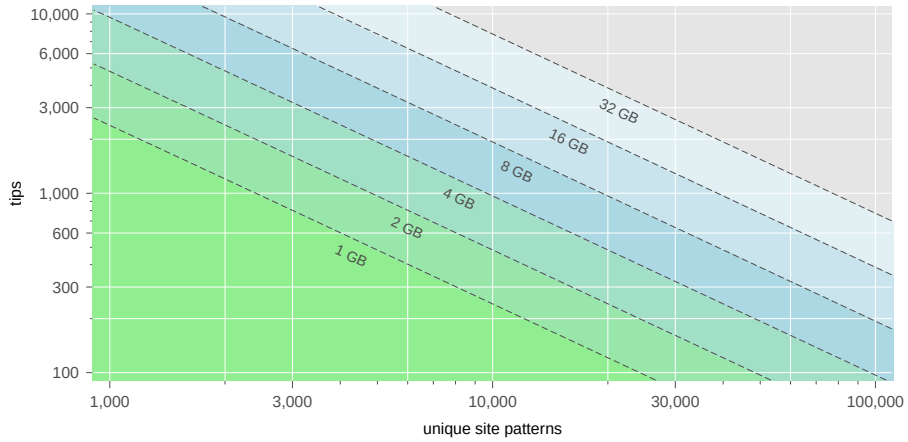


Figure 4 Log-log contour plot depicting BEAGLE-GPU memory usage for BEAST nucleotide model analyses with four-rate categories and double precision floating-point arithmetic, over a range of problem sizes in terms of number of tips and of unique site patterns. The amount of memory depicted as values below to dashed isolines convey the upper boundary for the memory size indicated. Memory requirements shown here assume an unpartitioned dataset. Partitioned analyses incur a small amount of memory overhead, typically less than 100 MB.

Figure 4 shows how much memory is required for problems of different sizes when running nucleotide model partitions in BEAST (Suchard et al., 2018) with BEAGLE GPU acceleration. Note that when multiple GPUs are available, BEAST can split the data into separate BEAGLE instances, one for each GPU. Thus each GPU will only require as much memory as necessary for the data subset assigned to it. Typical PC-gaming GPUs have 8 GB of memory or less, while GPUs dedicated to high performance computing, such as the NVIDIA Tesla series, currently have as much as 32 GB of memory.

For partitions with a discrete trait character, the memory required depends on the number of states the character can assume, in addition to the size of the tree (i.e., the number of tips). This memory requirement will typically be significantly less than that of the nucleotide data. As an example, for a tree with 1,000 tips, a discrete trait character partition with 100 possible states will use approximately 0.5 GB of GPU memory.

2.2.1 Hardware

Highly parallel computing technologies such as GPUs have overtaken traditional CPUs in peak performance potential and continue to advance at a faster pace. Additionally, the memory bandwidth available to the processor is especially relevant to data-intensive computations, such as the evaluation of nucleotide model likelihoods. In this measure as well, high-end GPUs significantly outperform equivalently positioned CPUs.

BEAGLE was designed to take advantage of this trend of increasingly advanced GPUs and uses runtime compilation methods to optimize code for which-ever generation of hardware is being used. For the analyses in the "Efficiently analysing large viral data sets in computational phylogenomics" chapter, we have used an NVIDIA Tesla P100 GPU, with 3584 CUDA cores and 16 GB of memory. Its peak figures for memory bandwidth and computational performance are of 720 GB/s and 4.7 trillion floating-point operations per seconds (TFLOPS). These figures are nearly an order of magnitude higher than those for a modern, high-end multi-core CPU. At the time of writing, the most powerful GPU for scientific computing is NVIDIA's Tesla V100, which comes equipped with 5120 CUDA cores and 32 GB of memory, with a memory bandwidth of 900 GB/s and a computational performance of up to 7.8 TFLOPS.

3 Discussion

BEAGLE is a high-performance computational library that offers substantial performance gains in phylogenetic and phylodynamic inference. Now at version 3, BEAGLE is fully integrated with BEAST 1.10.5 (Suchard et al., 2018) and MrBayes 3.2.7 (Ronquist et al., 2012), making use of the latest advances such as increased parallelism for nucleotide-model analyses on GPUs. We note that another high-performance library, known as the Phylogenetic Likelihood Library (Flouri et al., 2015), has been developed and integrated in two phylogenetic software packages: DPPDiv (Heath et al., 2011) and IQ-TREE (Nguyen et al., 2015). While benchmark tests (Flouri et al., 2015; Ayres et al., 2019) identify the strengths of both libraries in different scenarios, it is apparent that support for these libraries remains rather limited at the time of writing. We expect that with ever-growing data set sizes, both libraries will be increasingly adopted over time allowing parameter estimation for complex evolutionary models.

While offering substantial performance gains for statistical phylogenetics, the use of such high-performance libraries is not a panacea to enable complex model combinations on increasingly large data sets. Employing these libraries should be coupled with highly-efficient parameter estimation strategies, which have started to find their way into Bayesian phylogenetic and phylodynamic inference. To enable parallel estimation of a potentially large collection of continuous parameters, adaptive MCMC is able to exploit multi-core processing architectures to improve MCMC integration efficiency (Baele et al., 2017). Bayesian phylogenetics has also started adopting Hamiltonian Monte Carlo to improve inference efficiency of branch-specific evolutionary rates, by means of fast gradient evaluations (Ji et al., 2019). These efficient transition kernels have only recently seen their first implementations in phylogenetics and phylodynamics research, but offer promising avenues for further performance improvements.

References

Ayres, D. L. and Cummings, M. P. (2017a). Configuring concurrent computation of phylogenetic partial likelihoods: Accelerating analyses using the BEAGLE library. In Ibrahim, S., Choo, K., Yan, Z., and Pedrycz, W., editors, *Algorithms and Architectures for Parallel Processing. ICA3PP 2017, Helsinki, Finland*, volume 10393 of *Lect. Notes Comput. Sc.*, pages 533–547.

Ayres, D. L. and Cummings, M. P. (2017b). Heterogeneous hardware support in BEAGLE, a high-performance computing library for statistical phylogenetics. In *46th International*

Conference on Parallel Processing Workshops (ICPPW 2017), pages 23–32, Bristol, United Kingdom.

Ayres, D. L., Cummings, M. P., Baele, G., Darling, A. E., Lewis, P. O., Swofford, D. L., Huelsenbeck, J. P., Lemey, P., Rambaut, A., and Suchard, M. A. (2019). BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst. Biol.*, 68(6):1052–1061.

Baele, G., Lemey, P., Rambaut, A., and Suchard, M. A. (2017). Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST. *Bioinformatics*, 33(12):1798–1805.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019). Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4):1–28.

Flouri, T., Izquierdo-Carrasco, F., Darriba, D., Aberer, A., Nguyen, L.-T., Minh, B., Haeseler, A. V., and Stamatakis, A. (2015). The Phylogenetic Likelihood Library. *Syst. Biol.*, 2(1):356—362.

Guindon, S., Dufayard, J.-F. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3).

Heath, T. A., Holder, M. T., and Huelsenbeck, J. P. (2011). A Dirichlet Process Prior for Estimating Lineage-Specific Substitution Rates. *Molecular Biology and Evolution*, 29(3):939–955.

Ji, X., Zhang, Z., Holbrook, A., Nishimura, A., Baele, G., Rambaut, A., Lemey, P., and Suchard, M. A. (2019). Gradients do grow on trees: a linear-time O(N)-dimensional gradient for statistical phylogenetics. https://arxiv.org/pdf/1905.12146.pdf.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542.

Suchard, M., Lemey, P., Baele, G., Ayres, D., Drummond, A., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):vey016.

Suchard, M. A. and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics*, 25:1370–1376.

Swofford, D. L. (2003). *Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* Sinauer Associates, Sunderland, Massachusetts.

Zwickl, D. J. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.* PhD thesis, The University of Texas at Austin.

Chapter 5.5 Species Delimitation

Bruce Rannala
Department of Evolution and Ecology, University of California Davis
One Shields Avenue, Davis CA USA
brannala@ucdavis.edu
https://orcid.org/0000-0002-8355-9955

Ziheng Yang[1]
Department of Genetics, Evolution and Environment, University College London
London WC1E 6BT, United Kingdom
z.yang@ucl.ac.uk
https://orcid.org/0000-0003-3351-7981

Abstract

Species delimitation is the process of determining which groups of individual organisms constitute different populations of a single species and which constitute different species. The problem goes back to the earliest days of taxonomy and formalized processes for describing new species exist and are widely used, although the methods are time-intensive and problematic for some species. Genomic data carries extensive information about the degree of genetic isolation among species and about ancient and recent introgression. For this reason, genomic data can play an important role in species delimitation under many existing species concepts. Here we review the history of molecular species delimitation leading up to the current genomic era. We then describe the most widely used computational methods for species delimitation using single- and multi-locus genomic data. Relative strengths and weaknesses of the approaches are discussed and a new method for delimiting species based on empirical criteria is proposed.

How to cite: Bruce Rannala and Ziheng Yang (2020). Species Delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 5.5, pp. 5.5:1–5.5:18. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 What is Species Delimitation?

Species play a central role in all branches of biological research and are the fundamental unit to measure biodiversity. The current rate of extinction of species on the planet due to anthropogenic activity is difficult to precisely estimate both because species boundaries are in some cases unclear and because millions of species have yet to be described. it is estimated that 80 to 90% of species on planet Earth are undiscovered, and it is likely that numerous contemporary species have already become extinct without scientists ever having documented their existence. Species delimitation is therefore an activity central to conservation of biodiversity. Several large initiatives are underway to either barcode most species (for example, by sequencing a single locus for millions of species) (International Barcode of Life `http://ibol.org/`), or sequencing whole genomes of all known species and discovering the remaining species (Earth Biogenome Project `https://www.earthbiogenome.org/`). Taxonomists are presently in a race to document the species in many groups that are

[1] Z.Y. is supported by a Biotechnological and Biological Sciences Research Council grant (BB/P006493/1).

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 5.5; pp. 5.5:1–5.5:18

A book completely handled by researchers.
No publisher has been paid.

en route to extinction. All these efforts require methods for determining which individuals constitute new species (species delimitation) or should instead be assigned to an existing species (species assignment). Genomic species delimitation, the topic of this chapter, is therefore at the forefront of modern biodiversity science.

Taxonomic science has at least three distinct, but interdependent, roles in biology: the assignment of individual organisms into pre-existing species categories (species assignment), the assignment of species to higher categories (genus, family, etc), and the designation of new species categories to accommodate individual organisms that do not fit into an existing species category (species delimitation). Historically, all three roles were performed by taxonomists using morphological characters. For many species this has been effective and uncontroversial. However, some domains of life, such as bacteria, have few distinctive traits while others have distinctive morphologies that are highly plastic and associated with environmental factors leading to morphological convergence or divergence that is incompletely associated with genetic or evolutionary relatedness. For such groups, morphological delimitation of species can sometimes fail dramatically. Moreover, morphological species delimitation requires a high level of expertise and is time consuming, often making it impractical for large groups undergoing extinctions, which urgently need to be classified. A semi-automated process of delimitation in which the role of experts is to verify and refine results obtained from genomic data and computer algorithms is therefore very attractive. Such developments do not obviate the need for taxonomists but may alter their role in the process of documenting biodiversity. Automated algorithmic methods for delimiting species using genome data are the subject of this chapter. As noted by Jörger and Schrödl (2013) taxonomy includes both species discovery (delimitation) and the subsequent process of establishing a formal diagnosis and naming scheme. Diagnoses based on DNA are not considered here. As noted by Sites Jr and Marshall (2004), species concepts have received much more study than have "operational criteria" to be used for defining species in empirical studies. Operational criteria for use with molecular sequence data are the sole focus of this chapter.

2 A Brief History

2.1 Numerical taxonomy

The concept of using computer algorithms to delimit species based on morphological characters traces back to the 1960s. With the rise of computers and multivariate statistical methods the new field of Numerical Taxonomy (Sneath, 1957a) appeared poised to offered an objective solution to many of the problems of assigning and classifying species with confusing morphological variation. Bacterial taxonomists were early to adopt numerical taxonomy (Sneath, 1957b), possibly in the hope that large numbers of trait measurements could compensate for a lack of distinctive traits among bacterial strains (Goodfellow, 1971). Sneath (1976) considered a model of phenetic variation in bacteria, for example, in which the phenotype of a species is represented as a spherical multivariate normal distribution with the quantiles defining the species boundaries. Numerical taxonomy did not eliminate problems such as convergent evolution, however, which could cause unstable classifications based on morphology. Moreover, while multivariate analyses of morphology were effective at clustering individual operational taxonomic units (OTUs) into groups that shared distinctive features they did not provide *a priori* means for establishing species boundaries and could not identify the sources of morphological variation (genetic versus environmental).

2.2 Molecular taxonomy

During the 1970s, as molecular sequence data became available for proteins, and later for DNA and RNA, their potential utility for classifying difficult groups such as bacteria was widely recognized (Fox et al., 1977; Wayne et al., 1987; Wilson, 1995). During the 1980s, DNA-DNA hybridization (DDH) became the taxonomic gold standard for diagnosing new species of Bacteria and Archaea with a DDH similarity of less than 70% considered evidence of distinct species (Wayne et al., 1987; Meier-Kolthoff et al., 2013). However, for most groups of organisms the role that molecular versus morphological data should play in species assignment and delimitation – as well as the diagnostic criteria to be applied to each data type – remained uncertain. Traditional morphological delimitations rely on the identification of fixed differences between species. However, even for morphological characters finding convincing evidence of fixed differences requires sample sizes much larger than those found in most studies (Wiens and Servedio, 2000). Moreover, for molecular data with thousands (or millions) of bases in a sequence the probability of finding spurious fixed differences in a small sample of individuals can be very high. Fixed differences of bases at individual sites, or even of haplotypes, are commonly observed among populations within a species and are therefore not a sufficient criterion for delimiting species. Quantitative delimitation approaches that used diploid genotypic molecular data (such as allozymes) and population genetic statistics were proposed by several groups (reviewed by Sites Jr and Marshall, 2004). In particular, measures of gene flow based on Wright's F_{st} statistics were proposed as a criterion for delimiting species, with populations grouped into a single species if gene flow is detected (Porter, 1990). Criteria based on gene flow are unsatisfactory because they depend on overly simple models of population structure and a subjective threshold of gene flow for delimiting species. Furthermore, comparative genomic analyses in the past decade have highlighted the fact that gene flow between species is common in both plants and animals (Ellegren et al., 2012; Wu et al., 2018), including humans and their close relatives (Nielsen et al., 2017).

2.3 Patterns in gene trees

The widespread use of multilocus sequence data in the 1990s, and the development of the coalescent theory in population genetics, led several groups to develop species delimitation criteria based on observed patterns in gene trees. Avise and Ball Jr (1990) proposed a "genealogical species concept" that led to the development of several operational definitions based on patterns of shared ancestry in inferred gene trees. The most widely used criterion was "exclusivity" , also referred to as "reciprocal monophyly" (Ball et al., 1990; Baum and Donoghue, 1995; Palumbi et al., 2007), which means that all sequences from one species form a monophyletic group in the gene tree relative to the sequences from any other species. Such methods treat gene trees as observations and do no properly account for errors in inferred gene trees, which tend to reduce the frequency of reciprocal monophyly. This criterion may be too strict in some situations and not strict enough in others, depending on the population isolation time relative to the population sizes. If the populations are very large the isolation time required to reach exclusivity at all loci (or at 50% of loci, to use a weaker criterion of exclusivity) can be very long (Hudson and Coyne, 2002; Hudson and Turelli, 2003; Knowles and Carstens, 2007), and the exclusivity criterion may fail to recognise good species. In contrast, if the populations are very small (as in the case of a few founders establishing a population), the gene tree may be reciprocally monophyletic, but the population isolation time may be too short to consider them as distinct species. Templeton (2001) proposed a

genealogical "test of cohesion" for identifying species using his method of nested clade analysis (NCA). However, NCA is known to have poor statistical properties as it does not account adequately for demographic stochasticity and has extreme type I error rates (Beaumont and Panchal, 2008; Petit, 2008; Knowles, 2008).

2.4 DNA barcoding

Debate concerning procedures for species assignment and delimitation using molecular data was reinvigorated during the 2000s with the publication of an influential proposal for a "DNA barcoding" initiative using a single locus (the COII mitochondrial gene for animals) as a diagnostic for assigning species (Hebert et al., 2003). Advocates of DNA barcoding also proposed the use of sequence divergence thresholds (or barcode gaps) as a means for delimiting new species (Tautz et al., 2003; Blaxter, 2003). Criticisms of species delimitation using DNA barcoding included the fact that interspecific and intraspecific sequence distances may be similar in large populations, so that no fixed threshold exists, and that thresholds are subjective and often difficult to establish *a priori* (Moritz and Cicero, 2004; Will and Rubinoff, 2004; DeSalle et al., 2005). Methods based on a single locus are also expected to have low power for many recently diverged species. An advantage of using a single diagnostic locus was that it enabled high throughput analyses not possible with the multilocus sequencing approaches of the 2000s. The advance of sequencing technologies has shifted recent interests to multilocus species delimitation and assignment methods, which are more powerful than single locus methods.

2.5 Multispecies coalescent

The barcoding debate prompted several researchers to point out that the multispecies coalescent (MSC) (Rannala and Yang 2003; Chapter 3.3 [Rannala et al. 2020]) could provide a probability distribution for the likely gene trees given a particular species tree (in the absence of gene flow between species) and that this could provide a statistical model for assigning individuals to species based on either single-locus (Pons et al., 2006) or multilocus (Nielsen and Matz, 2006) sequence data. These approaches also offered an alternative to barcoding-inspired delimitation methods based on percent sequence divergence between species (Hebert et al., 2003) which is sensitive to the levels of polymorphism within populations. Methods based on the multispecies coalescent do not require reciprocal monophyly to delimit species (Knowles and Carstens, 2007) and can take proper account of statistical uncertainty in gene trees (Yang and Rannala, 2010). However, early methods were not able to analyze multilocus datasets and used approximations to the MSC (Pons et al., 2006). More recently, several approximate and exact multilocus methods have been developed for species delimitation under the MSC (O'Meara, 2009; Yang and Rannala, 2010) that can potentially scale to genomic datasets and provide greater power for species delimitation and species assignment. Inferences from such methods may be sensitive to model violations, however, such as gene flow between species (Leaché et al., 2018). MSC-based methods are an example of a parametric inference method.

2.6 Machine learning

The high performance of machine learning algorithms in solving many classification problems, and the straightforward generic nature of their application, has led to many recent applications in population genetics, phylogenetics, and other areas of evolutionary biology. Machine

learning algorithms can be broadly classified as either supervised (SML) or unsupervised (UML) machine learning, according to how training datasets are utilized. Both SML and UML approaches have been recently applied to species delimitation (Pei et al., 2018; Derkarabetian et al., 2019). Pei et al. (2018) developed an SML algorithm for species delimitation using support vector machines. Datasets generated by population genetic simulations (with or without gene flow) were used to train the algorithm. Summaries of the data were used rather than using sequence data directly to make the model and computation tractable. Five summary statistics were used: the proportion of private positions, the folded site frequency spectrum, the pairwise difference ratio, F-statistics, and the longest shared tract. For simulated data, the algorithms appeared to perform as well as the model-based species delimitation method BPP (Yang and Rannala, 2010) when species are genetically isolated and were more likely than BPP to delimit species that experienced gene flow. A similar approach was developed by Smith and Carstens (2019) but using the folded site frequency spectrum and a Random Forest (RF) classifier.

SML methods have the advantage that they can be computationally efficient and can be trained on models that are too complex to derive formal Bayesian or maximum likelihood estimators. A well-known weakness of supervised learning methods, encapsulated by the so-called "supervised learning no free lunch theorem" (Wolpert, 2002), is that they can become too specialized – they work very well for the training dataset but poorly for many other datasets. In this case, since the algorithm is trained using specific values of population genetic parameters such as θ and M it could perform poorly outside the training range. Formal statistical methods do not have this problem since they have optimality properties that hold over the entire parameter space. For complicated models it may be impossible to train an SML algorithm over the entire state space for the parameters. Another weakness of both SML and UML methods is that they often use summary statistics rather than the full dataset. Unless the summary statistics are known to be sufficient statistics this will entail loss of information. Finally, little is known about the asymptotic statistical performance of most SML or UML methods and developers must therefore resort to simulations to evaluate them when inferring evolutionary parameters for which the true parameter values are unknown. Simulation studies can never be comprehensive.

3 A Survey of Species Delimitation Methods

Here we provide a concise survey of the most widely used species delimitation methods, sketching important features of the statistical and computational theory and the assumptions underlying them. We focus exclusively on methods designed for use with DNA sequence data and divide the methods into two categories: (1) heuristic methods, which use a summary statistic or algorithm for delimitation that is not derived from a formal statistical model of the population genetic structure. Heuristic methods are often computationally efficient but the results can be difficult to interpret and they may have poor statistical properties; (2) parametric methods, which are based on an explicit probabilistic model of population divergences and evolution of the genetic sequences and which select the delimitation model that maximizes the likelihood or Bayesian posterior probability. Full-likelihood methods under a parametric model are known to be asymptotically most powerful when the model is correct but are often computationally demanding. If the model is incorrect the statistical properties may become unpredictable. Most widely used parametric methods are derived based on the MSC model, which will therefore be described in some detail below. Both approximate and exact parametric methods will be described. A distinction is made between methods

designed for use with a single locus versus multilocus sequence data. We focus on methods that are applicable across the tree of life, and do not consider methods developed specifically for a certain species group, such as Genome Blast Distance Phylogeny (Meier-Kolthoff et al., 2013) for species delimitation of bacteria.

3.1 Heuristic methods

Most heuristic methods for species delimitation originate from the DNA barcoding initiative and are therefore designed for single locus data. Multilocus heuristic methods proposed more recently often simply concatenate genes (Zhang et al., 2013). Heuristic methods are computationally efficient and can be applied to large datasets. Their statistical performance may be good in some regions of the parameter space but poor in others. To evaluate the performance, simulation is often used because standard statistical theory (e.g., asymptotic efficiency in large datasets, etc) typically does not apply to heuristic methods.

3.1.1 ABGD: Automated Barcode Gap Discovery

Early studies of pairwise sequence divergence between and within species aimed to identify a "barcode gap" which can distinguish within-population differences from differences caused by species divergences. For example, a relative difference of one order of magnitude (the 10× rule) between intra- and interspecific divergences was proposed by Hebert et al. (2004) and a maximum of 3% for intraspecific divergence (the 3% rule) was proposed by Smith et al. (2005). Such *a priori* thresholds are arbitrary and subsequent analyses suggested that the barcode gap varies among species groups based on factors such as effective population sizes and species divergence times (Hickerson et al., 2006; Rannala, 2015).

The aim of the automated barcode gap discovery (ABGD) program (Puillandre et al., 2012) is to identify barcode gap thresholds in an automated process. For a sample of n haploid sequences all $n(n-1)/2$ pairwise distances are calculated. The distance metric may use a correction for multiple-hit substitutions. The distances are then ranked in increasing order so that $d_i \leq d_{i+1}$ for $i = 1, \ldots, n(n-1)/2$. For the distance of rank r the local slope is calculated as

$$s_{r,w} = \frac{d_{r+w} - d_w}{w}. \tag{1}$$

The slope is expected to be largest at the barcode gap that delimits species. Thus the method infers the barcode gap as the distance that maximizes the local slope.

Simulation under the coalescent process was used to compute a threshold value under which sequences are more likely to be intraspecific, assuming a constant population size with n and θ specified and estimating the threshold exceeding 95% of pairwise distances. It was concluded that for $n > 10$ the threshold was a linear function of θ with a slope of 2.581. The rationale was that, if θ is known, the barcode gap can be predicted from the simulation results. The parameter θ is unknown for empirical data but for a single population the average pairwise distance provides an estimate of θ. With more than one species in the sample the estimate of θ will be too large. A user-specified "prior" threshold is therefore used to separate intraspecific distances from interspecific distances. Only putative intraspecific distances (those below the prior threshold) are used in estimating θ. Once θ is estimated a threshold distance is determined, and groups are then formed so that "the distance between sequences from different groups is always larger than the gap distance, and for each sequence of each group, there is at least one other sequence in the group at a distance smaller than

the gap distance." This procedure is applied recursively to identify additional groups that may have different barcode gaps.

Like the PTP method discussed below, ABGD may be expected to work best when the gene tree is essentially monophyletic (i.e., there is no incomplete lineage sorting). Like other barcoding methods it only uses information from a single locus. It is computationally inexpensive and can be applied to large samples of sequences. Its weaknesses include its reliance on simple pairwise distance calculations and clustering operations, and failure to use all the information in the sequence data. The distribution of intraspecific distances may be multimodal due to factors such as population growth or selection (Rogers and Harpending, 1992; Harpending et al., 1993) potentially leading to spurious barcode gaps.

3.1.2 GMYC: General Mixed Yule Coalescent

The General Mixed Yule Coalescent (GMYC) (Pons et al., 2006) assumes that the waiting times between coalescence events or branch lengths in a gene tree fall into two classes: those within species with the rate determined by the coalescent process, or those between species with the rate determined by a generalization of the Yule process model of species divergences. It is assumed that a time point T exists at which the generative process for gene-tree nodes switches from a coalescent process to a Yule process. Maximum likelihood is used to estimate T and the choice of T determines the species delimitation. A likelihood ratio test is also proposed to test for the existence of multiple species ($T > 0$) versus a single species ($T = 0$). This method treats the inferred gene tree as known and is therefore computationally simple and fast, but has the drawback of ignoring errors in the gene tree. Another limitation is that the method can be applied to only a single locus, although an attempt was made to extend it to multiple loci (Fujisawa and Barraclough, 2013). More importantly, by using a single threshold T, the model implicitly assumes that all lineages in each population coalesce before any speciation event occurs, implying the absence of incomplete lineage sorting. The method ignores the coalescent process within ancestral species populations, in effect assuming that the gene tree is reciprocally monophyletic. The GMYC method should thus perform best for datasets with long intervals between speciation events and small population sizes, in which case incomplete lineage sorting is unlikely to occur.

3.1.3 PTP: Poisson Tree Process

The "Poisson Tree Process" (PTP) identifies species status based on the distribution of branch lengths in the gene tree (Zhang et al., 2013). The tree and branch lengths are inferred from a sequence alignment using maximum likelihood and then treated as known without errors. When multiple loci are available they are concatenated to infer one gene tree with branch lengths. The PTP method models the branch lengths, x_i, in a rooted non-ultrametric tree as a mixture of two exponential distributions with rates λ_S (between-species) and λ_C (within-species), respectively. A species delimitation model (Λ) assigns every branch on the gene tree to one of those two classes. The log-likelihood for the delimitation, given the data (G) of a rooted gene tree with n branches, is then

$$L(\Lambda; G) = \sum_{i=1}^{k} \log(\lambda_S e^{-\lambda_S x_i}) + \sum_{i=k+1}^{n} \log(\lambda_C e^{-\lambda_C x_i}), \tag{2}$$

where the delimitation model Λ partitions k branches as "between-species" and $n-k$ branches as "within-species" . The Poisson rates (λs) are estimated by using the inverse of the average

branch length in each class, so that no iteration is needed for parameter estimation. A heuristic search is used to find the delimitation that maximizes the likelihood. An extension of the method (Kapli et al., 2017) allows λs to vary among populations.

A strength of the PTP approach is that it can handle large datasets with thousands of species. Unlike the GMYC method discussed above, PTP uses a rooted non-ultrametric tree so that it does not rely on the molecular clock, which may be seriously violated for distantly related species. The approach implicitly assumes reciprocal monophyly in the gene tree and a perfect match of the gene tree with the species tree. It is thus expected to work best for identifying species that are separated by long intervals between speciation events and that have small population sizes.

Some weaknesses of the method may be easily identified. The method is essentially a single-locus method, as concatenating sequences across gene loci or genomic segments does not account for the stochastic fluctuations in the coalescent process among the loci. Because the Poisson tree process describes a distribution of delimitation models, in theory there should be a prior term $f(\Lambda)$ in equation 2, and the posterior probability for the delimitation model should be maximized $f(\Lambda|G)$ instead of the log-likelihood:

$$f(\Lambda|G) \sim f(\Lambda) \prod_{i=b}^{n} \lambda_{\mathbb{I}_b} \exp\{-\lambda_{\mathbb{I}_b} x_i\} \tag{3}$$

where the indicator $\mathbb{I}_b$ = 'S' or 'C' depending on whether delimitation model Λ partitions branch i as a between-species branch or a within-species coalescent branch. Furthermore, if more than two sequences are in one species according to the delimitation model, the waiting time until the next coalescent when there are k lineages is proportional to $2/(k(k-1))$. This can in theory be accommodated by redefining λ_C as the coalescent rate for two lineages, and applying the correction factor $2/(k(k-1))$ if the within-species branch represents the waiting time until the next coalescent when there are k lineages in the sample in the population. Parameters λ_S and λ_C may be estimated using the method of moments, by using the average observed values so that the amount of computation remains the same. While it is unnecessary to sample multiple individuals or sequences to infer the species phylogeny, sampling multiple sequences from the same species adds much information in species delimitation (Zhang et al., 2011). Such modifications may make it possible to utilize the information in multiple samples.

3.2 Parametric methods

Parametric species delimitation methods are based on a probabilistic data-generating model, that is, a model of the biological processes generating gene genealogical trees and DNA sequences among individuals. This entails modeling the process of speciation, the genealogical process of coalescence within populations, and the process of DNA sequence evolution driven by genetic drift and natural selection. The canonical model relevant to species delimitation is the multispecies coalescent (MSC) with neutral evolution (Rannala and Yang, 2003). All parametric species delimitation methods considered here are based on an MSC model, although some methods (such as GMYC, Pons et al., 2006) are based on simplifications or approximations of it. The MSC model, describing the basic biological processes of reproduction and drift, is important to heuristic methods of species delimitation as well: for example, evaluation of the performance of those methods typically involves simulating genetic sequence data under the MSC model. If species delimitation can be formulated as a problem of statistical inference under the MSC model, standard theories of statistical inference can be made use of; for example, maximum likelihood and Bayesian methods are known to have

desirable asymptotic (large-sample) properties such as consistency and efficiency. However, the assumptions of the parametric approach may be violated in real data analysis, and an important question is how well the method performs when the model is misspecified.

3.2.1 Multispecies coalescent and the distribution of gene trees

The genomes of individual organisms carry rich historical information at many levels and at different time scales. At a low level, genomes provide information about inbreeding (homozygosity), pedigree relatedness and genome sharing (identity by descent) among individuals. At a higher level, genomes are informative about population or species affiliations and the evolutionary relationships among species. Different statistical models and inference methods are often used to extract information at these different levels, yet the models used are interrelated and have many shared parameters. For example, individual inbreeding coefficients are determined by mating patterns and population structure, and gene trees of relationships among sequences are influenced by population level processes such as selection and genetic drift. Phylogenetic trees among species are, in turn, inferred from patterns observed in gene trees.

The phylogenetic relationships of species or genetically isolated populations constrain the possible relationships of genomic sequences (see Chapter 3.3 [Rannala et al. 2020]). When we trace the genealogical history of a sample of sequences backwards in time, sequences from different isolated populations cannot coalesce until we reach the common ancestor of those populations. If the sequences are in the same population the rate at which they coalesce is determined by the population size. Thus, the phylogenetic tree of populations generates a probability distribution on the possible gene trees. Conversely, the probabilities of gene trees, which underlie the genomic sequence data, can inform us about population history, such as population divergence times, population sizes, and between-population migration or introgression.

Parametric statistical approaches to species delimitation use the multispecies coalescent (MSC) model to infer the existence of genetically isolated populations that are potential species. Multilocus sequence data sampled from modern species or populations are used to calculate the posterior probabilities (Yang and Rannala, 2010) or marginal likelihoods (Grummer et al., 2013; Rannala and Yang, 2017) of different species delimitation models, where a delimitation model corresponds to a certain way of merging populations into the same species (Figure 1). In a randomly mating population the two sequences sampled from a diploid individual are equivalent to any two sequences randomly sampled from the population so only the population identity of each sequence is relevant. If strong evidence exists that two or more populations are genetically isolated, they will be assigned to distinct species. In contrast, if they constitute a single panmictic population they will be collapsed into a single species. In this approach, the level of gene flow may have a major impact, as will the amount of time since population divergence. The procedure does not derive from any particular species definition but it includes the Biological Species Concept as a particular case, so that the species status is recognized given a sufficient period of reproductive isolation.

The MSC model is parametrized by the species tree topology, as well as the species divergence times (τs) and population sizes for both modern and ancestral species (θs) (Figure 1). A (rooted) species tree for s species has $s-1$ internal nodes or speciation events and $2s-1$ nodes or populations; thus an MSC model for s species has $s-1$ node ages or species divergence times (τs) and $2s-1$ population size parameters (θs). In analysis of genomic sequence data, both τs and θs are measured by genetic distance or the expected number of mutations per site.

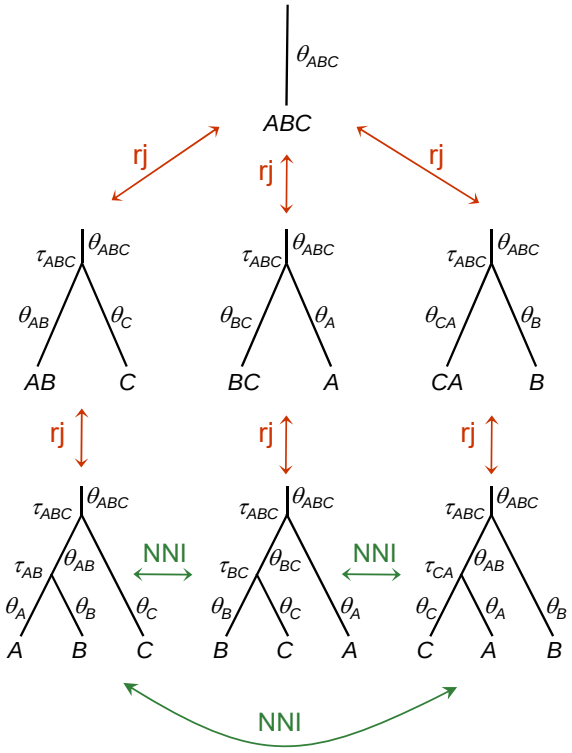


Figure 1 For three populations (A, B, C), there are five species delimitation models (one of 1 species, three of 2 species, and one of 3 species) and seven MSC models, with the delimitation of three species (bottom row) resolved into three MSC models (three species phylogenies). Each MSC model of s species has $(s-1)$ divergence times (τs) and $(2s-1)$ population size parameters (θs). Species delimitation through Bayesian model selection uses Markov chain Monte Carlo (MCMC) proposals to traverse the space of MSC models to estimate the posterior probabilities for the MSC models. The posterior for a delimitation model is the sum of the posterior probabilities for the compatible MSC models: for the example here, the probability for the existence of three species (A, B, C) is the sum of the three MSC models on the bottom row. In BPP, subtree-prunning-and-regrafting (SPR) or nearest-neighbor-interchange (NNI) algorithms are used to move between different species phylogenies, and rjMCMC is used to move between different species delimitations (Yang and Rannala, 2010, 2014).

3.2.2 Posterior probabilities of species delimitation models

Given a set of K populations, different species delimitations correspond to different ways of merging populations into the same species, with the number of delimited species ranging from 1 to K. We assume that individuals are correctly assigned to populations, and a population will not be split into different species although multiple populations may be merged into the same species. As an extreme approach, each sampled individual can be assigned its own population, and the Bayesian algorithm can be used to achieve both assignment and delimitation (Olave et al., 2014). When more than two species exist in the delimitation model, there are different species phylogenies as well. The species delimitation and species phylogeny together constitutes a fully specified MSC model, which allows the definition of the parameters and the specification of the probability distribution of the gene trees (Rannala and Yang, 2003). For example, with 3 populations, there will be five delimitation models: one model of 1 species, three models of 2 species, and one model of 3 species (Figure 1),

and in the case of 3 species, there are also three species phylogenies. Thus in total there are seven MSC models.

Let $\boldsymbol{\theta}_k$ be the parameters in the MSC model specified by a delimitation model Λ_k. Note that the delimitation alone (e.g., knowledge of the existence of three species without knowledge of the species phylogeny) may be insufficient to define the parameters or to specify the data-generating mechanism. Thus from now on we assume that the delimitation model Λ_k also specifies the species phylogeny so that the parameters can be defined. In the case of figure 1 for three populations, there are seven such models. The posterior probability of delimitation model Λ_k given the sequence data at L loci, $\boldsymbol{X} = \{X_i\}$, is then

$$\mathbb{P}\{\Lambda_k|\boldsymbol{X}\} \propto \pi_k M_k, \tag{4}$$

where π_k is the prior probability for model Λ_k, and M_k is the marginal likelihood for model Λ_k. The proportionality constant is to ensure that the posterior probabilities for all models sum to 1. The marginal likelihood M_k for delimitation model Λ_k is an integral (an average) over all possible gene trees at each locus and over the MSC parameters,

$$M_k = \iint f(\boldsymbol{\theta}_k|\Lambda_k) \prod_i^L [f(G_i|, \Lambda_k, \boldsymbol{\theta}_k) f(X_i|G_i)] \, \mathrm{d}G \, \mathrm{d}\boldsymbol{\theta}_k, \tag{5}$$

where $f(\boldsymbol{\theta}_k|\Lambda_k)$ is the prior on the parameters under the model, $f(G_i|, \Lambda_k, \boldsymbol{\theta}_k)$ is the MSC density for gene tree G_i at locus i (Rannala and Yang, 2003), and $f(X_i|G_i)$ is the probability of the sequence alignment at locus i, known as the phylogenetic likelihood (Felsenstein, 1981).

The integrals of equation 5 are typically calculated numerically in the Markov chain Monte Carlo (MCMC) algorithm, as implemented in the BPP program (Yang and Rannala 2010; Rannala and Yang 2017; Chapter 5.6 [Flouri et al. 2020]). In other words, MCMC is used to traverse the model space, and the frequency at which the MCMC visits each model is the estimate of the posterior probability for that model. If a delimitation is compatible with multiple MSC models (for example, the delimitation of three species is resolved into three MSC models or three species phylogenies in Figure 1), its posterior probability is calculated as the sum of the posterior probabilities for the compatible MSC models. This is the A11 analysis by Yang (2015). In BPP, tree perturbation algorithms such as nearest-neighbor-interchange (NNI), subtree-pruning-and-regrafting (SPR), and NodeSlider are used to propose moves from one species tree to another with the species delimitation fixed (Yang and Rannala, 2010, 2014; Rannala and Yang, 2017). Moves between species delimitation models involve changes of dimension (the number of parameters), so they are implemented using a pair of reversible-jump MCMC moves ("split" and "join" , Yang and Rannala, 2010). For example, the "split" move can be used to move from the 2-species model (AB, C), which has 4 parameters, to the three-species model $((AB)C)$, which has 7 parameters, with three new parameters $(\tau_{AB}, \theta_A, \theta_B)$ created. The reverse "join" move changes the 3-species model to the 2-species model, dropping the redundant parameters. The pair of moves constitutes one reversible-jump proposal. RjMCMC algorithms often mix poorly, but thanks to development of improved algorithms, which make coordinated changes to the species tree and the gene trees when the MSC model changes (Rannala and Yang, 2013, 2017), BPP can be used to analyze datasets with hundreds or even thousands of loci. Simulation has confirmed the overall efficiency of the method (Zhang et al., 2011, 2014).

A number of empirical studies have found that the approach to model selection implemented in BPP tends to "oversplit" , favouring delimitation models with a large number of species (Sukumaran and Knowles, 2017). In some studies, the delimited number of species is the number of populations in the dataset. The problem appears to be worse when more

data (more loci) are analyzed. Leaché et al. (2018) studied the dynamics of Bayesian model selection when the true model involves two species/populations with migration but the two compared models assume either one species or two species without gene flow. The two models considered by BPP are in this case both wrong. However, analysis suggest that the two-species model is closer (judged by Kullback-Leibler divergence) to the true model of two populations with migration and is thus less wrong than the one-species model (Leaché et al., 2019). In such a case, the two-species model will dominate, with its posterior probability approaching 100% when the amount of data (the number of loci) approaches infinity. This provides an explanation for the observation that BPP tends to favour the model of distinct species even if there is substantial gene flow between the populations. If Bayesian model selection is conducted under the MSC model with migration or introgression, the two-species model will be the correct one and will naturally dominate, and the problem of over-splitting will become even more serious. With gene flow, the distinction between populations and species in such models is arbitrary. We suggest that the approach of Bayesian model selection be used mostly in the context of delimiting sympatric species whose distinctness is maintained by a lack of gene flow rather than partial genetic isolation.

3.2.3 Bayes factor delimitations

Several groups (Grummer et al., 2013; Leaché et al., 2014) have suggested the use of Bayes factors to choose among a small set of species delimitation models. The Bayes factor for two delimitation models Λ_1 and Λ_2 is defined as the ratio of the marginal likelihoods under the two models. From equation 4,

$$BF_{12} = \frac{M_1}{M_2} = \frac{\mathbb{P}\{\Lambda_1|X\}/\mathbb{P}\{\Lambda_2|X\}}{\pi_1/\pi_2}. \tag{6}$$

In other words, the Bayes factor is the ratio of the posterior odds to the prior odds. If we assign uniform prior probabilities $\pi_1 = \pi_2$ to the two models, the Bayes factor will simply be the posterior odds. Note that here the delimitation models Λ_1 and Λ_2 should be fully specified MSC models: if there are 3 or more delimited species, the species phylogeny should be considered part of the model specification as well. Otherwise the marginal likelihood is not well defined. Second, the marginal likelihood is an average over the prior distribution of parameters in the MSC model (the τs and θs). As a result, the prior on parameters may be influential on the marginal likelihood, besides the model of species divergences. The marginal likelihood is sometimes referred to as the "evidence" by Bayesians, but it should be borne in mind that it incorporates information from the prior which may represent subjective "opinions". Bayes factors and posterior probabilities for delimitation models are equivalent if Bayes factors are applied to the complete set of all possible delimitation models. Otherwise Bayes factors provide a local comparison among the delimitations examined. Interpretation or calibration of the Bayes factor is through reference to posterior model probabilities: a Bayes factor of 99 (= 0.99/0.01 in terms of posterior odds) might be considered "strong" evidence in favour of model Λ_1, for example, while 9 (= 0.9/0.1) might only be considered "positive" evidence.

Grummer et al. (2013) proposed a Bayes factor test of species delimitation and used the StarBeast program (Heled and Drummond, 2009) to calculate marginal likelihoods under different delimitation models. Leaché et al. (2014) developed an efficient Bayes factor delimitation method using the snapp (Bryant et al., 2012) algorithm to calculate marginal likelihoods. snapp is developed for single nucleotide polymorphism (SNP) data from unlinked loci. As every site (every SNP) is assumed to have its own independent history, the gene trees

including coalescent times can be integrated analytically under the MSC, eliminating the need for computationally expensive integration via MCMC. SNAPP is thus computationally efficient.

A major weakness of SNAPP, or of using unlinked SNP data, is information loss and a resulting lack of power. This occurs for two reasons. First, correction for ascertainment bias leads to loss of information. SNPs are collected because the sites are polymorphic. This fact has to be accommodated in the inference method, and correction for such "ascertainment bias" in data collection may lead to serious loss of information. As an analogy, there are two data outcomes in a binomial experiment: "success" and "failure". If we filter out all the "failures", it will be important to correct for such ascertainment bias and furthermore very little information remains after such data filtering. In the case of the SNP data, the information loss may not be so severe, but the correction may be expected to have a major impact on inference, in particular on branch lengths in the species phylogeny. It seems that the number of constant sites removed or the number of sites separating the SNP loci may be useful for recovering some of the information lost. Second, use of essentially independent SNP sites means that not all parameters in the MSC model are identifiable and the power for comparing different delimitation models may be affected as well. Multi-locus sequence alignments allow one to tease apart the among-site variation within the same locus given the gene tree (due to the Poisson mutation process) from the among-loci variation in the gene trees (due to the stochastic fluctuation of the multispecies coalescent process, influenced by parameters such as ancestral population sizes). As a result, all parameters in the MSC model are identifiable given the multi-locus sequence data. In contrast, a single SNP can only identify a single bipartition in a gene tree and contains very little phylogenetic information. The two sources of variation are then confounded. As a result, not all parameters in the MSC model are identifiable given SNP data.

3.2.4 Delimitation based on empirical criteria

Suppose we are given a detailed description of the history of two allopatric populations, including their divergence time, population sizes, and the timing, directions and intensity of migration or introgression events. Can we then decide whether the two populations belong to the same species or are two distinct species? If the two populations are sympatric, reduction or absence of gene flow revealed by genomic data will be evidence for genetic isolation and for the existence of barriers to gene flow so that distinct species status can be established. However, if the species are allopatric, genetic differentiation may be due to an absence of opportunities for interbreeding rather than the existence of isolation mechanisms or adaptation to different ecological and environmental conditions. Delineation of species boundaries in such cases will be arbitrary. The neutral genome has the power to let us infer a detailed population divergence history, but may not be informative about ecological adaptation or reproductive isolation. Delimitation in such cases may make use of empirically established criteria, based on evolutionary parameters. The MSC model, either without or with introgression, can be used to infer the MSC parameters and these then used in combination with any empirical criteria for species delimitation.

One such criterion is the genealogical divergence index (*gdi*) (Jackson et al., 2017). Suppose one samples two sequences (a_1 and a_2) from population A and one sequence (b) from population B. If sequences a_1 and a_2 coalesce first, the gene tree will be $G_a = ((a_1, a_2), b)$. Under the MSC model without gene flow, the gene tree probability is

$$P_a = \mathbb{P}(G_a|\boldsymbol{\theta}) = 1 - \frac{2}{3}\mathrm{e}^{-2\tau/\theta_A}, \tag{7}$$

where $2\tau/\theta_A$ is the population divergence time in coalescent units (with one coalescent time unit to be $2N_A$ generations) and $e^{-2\tau/\theta_A}$ is the probability that sequences a_1 and a_2 do not coalesce before reaching the time of species divergence (τ) when we trace the genealogy backwards in time. P_a ranges from $\frac{1}{3}$ to 1. Jackson et al. (2017) rescaled P_a to form the *gdi* index

$$gdi = (3 \times P_a - 1)/2, \tag{8}$$

so that it ranges from 0 to 1.

Based on the meta-analysis of Pinho and Hey (2010), Jackson et al. (2017) suggested the rule of thumb that *gdi* values < 0.2 suggest a single species and *gdi* values > 0.7 suggest distinct species, while *gdi* values within the range indicate ambiguous delimitation. Those limits correspond to 0.47 and 0.8 for P_a, and, in the case of no migration, to 0.22 and 1.20 for the population divergence time in coalescent units.

The use of the *gdi* index as a metric for delimiting species has several drawbacks. First, when populations A and B have very different sizes, it is unclear which population size should be used. For example, one can use two sequences from B (b_1, b_2) and one sequence from A to define the probability for the gene tree $G_b = ((b_1, b_2), a)$, and its probability $P_b = \mathbb{P}(G_b|\boldsymbol{\theta})$. However, when θ_A and θ_B are very different, P_a and P_b may be very different, leading to the awkward situation where P_A suggests one species while P_b suggests two (Leaché et al., 2019). Second, small population sizes may cause the index to over-split. It may therefore be useful to include a minimum absolute divergence time measured in generations. Third the criterion often leads to indecision, but this may reflect the difficulty of species delimitation for allopatric populations. In light of those drawbacks, we suggest the following modifications for an operational concept for delimiting species under the MSC model. We consider two populations to be distinct species if $P_a > 0.5$, $P_b > 0.5$, $M = Nm < 0.1$, and the divergence time is more than 10^4 generations. Otherwise we declare one species if $P_a < 0.4, P_b < 0.4, M = Nm \geq 1$, and divergence is within 10^3 generations. Situations between those two extremes are undecided.

Different heuristic criteria may be designed to correspond to different species definitions. Given any empirical criterion, a hierarchical procedure can be devised to delimit species using multi-locus genetic sequence data (Leaché et al., 2019). First we estimate a population phylogeny under the MSC. This is used as a guide tree so that its topology is not changed further, while the nodes on the tree may represent either populations or species. We then attempt to merge the populations into the same species using the criterion, starting from the tips of the tree, moving towards the root, each time re-estimating the MSC parameters. An ancestral node on the guide tree is merged into one species only if its descendant nodes are already merged. The procedure ends when no populations can be unambiguously merged into one species. The procedure is applied to several simulated and real datasets in Leaché et al. (2019).

4 Conclusions

Genomic data provide a rich source of information concerning the evolutionary history of species and populations such as divergence times, population sizes, and migration/hybridisation intensity. For sympatric species, convincing evidence of genetic isolation may be enough for establishing species status. For populations from different geographic locations, the genetic isolation can be due to isolation by distance, and may have to be combined with other sources of evidence to justify species status. In complex cases such as a species ring, the

situation may be so complicated that delimiting species becomes an arbitrary exercise. The MSC provides the framework for estimating evolutionary parameters including migration rates or introgression intensity, which can be used to apply empirical criteria for delimiting species or to generate hypotheses of species status, to be integrated with other evidence such as morphology and ecology (Yang and Rannala, 2010). An explicit extension of BPP to allow a model including a morphological character (iBPP) has been implemented by Solís-Lemus et al. (2015).

This chapter should have made clear the fact that there exists no "magic bullet" method for species delimitation using genomic data. However, methods are gradually converging on this target and already provide many useful tools for preliminary delimitations, as well as providing a solid theoretical foundation for future developments.

References

Avise, J. C. and Ball Jr, R. M. (1990). Principles of genealogical concordance in species concepts and biological taxonomy. *Oxford Surveys in Evolutionary Biology*, 7:45–67.

Ball, R. M., Neigel, J. E., and Avise, J. C. (1990). Gene genealogies within the organismal pedigrees of random-mating populations. *Evolution*, 44:360.

Baum, D. A. and Donoghue, M. J. (1995). Choosing among alternative "phylogenetic" species concepts. *Systematic Botany*, 20:560.

Beaumont, M. A. and Panchal, M. (2008). On the validity of nested clade phylogeographical analysis. *Molecular Ecology*, 17:2563–2565.

Blaxter, M. (2003). Molecular systematics: counting angels with DNA. *Nature*, 421:122.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular biology and evolution*, 29(8):1917–1932.

Derkarabetian, S., Castillo, S., Koo, P. K., Ovchinnikov, S., and Hedin, M. (2019). A demonstration of unsupervised machine learning in species delimitation. *Molecular Phylogenetics and Evolution*, 139:106562.

DeSalle, R., Egan, M. G., and Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1905–1916.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backstrom, N., Kawakami, T., Kunstner, A., Makinen, H., Nadachowska-Brzyska, K., Qvarnstrom, A., Uebbing, S., and Wolf, J. B. W. (2012). The genomic landscape of species divergence in ficedula flycatchers. *Nature*, 491:756–760.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.

Flouri, T., Rannala, B., and Yang, Z. (2020). A tutorial on the use of bpp for species tree estimation and species delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.6, pages 5.6:1–5.6:16. No commercial publisher | Authors open access book.

Fox, G. E., Pechman, K. R., and Woese, C. R. (1977). Comparative cataloging of 16s ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *International Journal of Systematic and Evolutionary Microbiology*, 27(1):44–57.

Fujisawa, T. and Barraclough, T. G. (2013). Delimiting species using single-locus data and the generalized mixed yule coalescent approach: a revised method and evaluation on simulated data sets. *Syst. Biol.*, 62:707–724.

Goodfellow, M. (1971). Numerical taxonomy of some nocardioform bacteria. *Journal of General Microbiology*, 69:33–80.

Grummer, J. A., Bryson Jr, R. W., and Reeder, T. W. (2013). Species delimitation using Bayes factors: simulations and application to the sceloporus scalaris species group (squamata: Phrynosomatidae). *Systematic biology*, 63(2):119–133.

Harpending, H. C., Sherry, S. T., Rogers, A. R., and Stoneking, M. (1993). The genetic structure of ancient human populations. *Current Anthropology*, 34:483–496.

Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270:313–321.

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology*, 2:e312.

Heled, J. and Drummond, A. J. (2009). Bayesian inference of species trees from multilocus data. *Molecular biology and evolution*, 27(3):570–580.

Hickerson, M. J., Meyer, C. P., and Moritz, C. (2006). DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology*, 55:729–739.

Hudson, R. R. and Coyne, J. A. (2002). Mathematical consequences of the genealogical species concept. *Evolution*, 56:1557.

Hudson, R. R. and Turelli, M. (2003). Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution*, 57:182.

Jackson, N., Carstens, B., Morales, A., and B.C., O. (2017). Species delimitation with gene flow. *Syst. Biol.*, 66:799–812.

Jörger, K. M. and Schrödl, M. (2013). How to describe a cryptic species? practical challenges of molecular taxonomy. *Frontiers in Zoology*, 10:59.

Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A., and Flouri, T. (2017). Multi-rate poisson tree processes for single-locus species delimitation under maximum likelihood and markov chain monte carlo. *Bioinformatics*, 33(11):1630–1638.

Knowles, L. L. (2008). Why does a method that fails continue to be used? *Evolution*, 62:2713–2717.

Knowles, L. L. and Carstens, B. C. (2007). Delimiting species without monophyletic gene trees. *Systematic Biology*, 56:887–895.

Leaché, A. D., Fujita, M. K., Minin, V. N., and Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP data. *Systematic biology*, 63(4):534–542.

Leaché, A. D., Zhu, T., Rannala, B., and Yang, Z. (2018). The spectre of too many species. *Systematic Biology*, 68:168–181.

Leaché, A. D., Zhu, T., Rannala, B., and Yang, Z. (2019). The spectre of too many species. *Syst. Biol.*, 68(1):168–181.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, 14:60.

Moritz, C. and Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS Biology*, 2:e354.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, 541:302.

Nielsen, R. and Matz, M. (2006). Statistical approaches for DNA barcoding. *Systematic Biology*, 55:162–169.

Olave, M., Sola, E., and Knowles, L. L. (2014). Upstream analyses create problems with dna-based species delimitation. *Syst. Biol.*, 63(2):263–271.

O'Meara, B. C. (2009). New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology*, 59:59–73.

Palumbi, S. R., Cipriano, F., and Hare, M. P. (2007). Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. *Evolution*, 55:859–868.

Pei, J., Chu, C., Li, X., Lu, B., and Wu, Y. (2018). CLADES: a classification-based machine learning method for species delimitation from population genetic data. *Molecular Ecology Resources*, 18:1144–1156.

Petit, R. J. (2008). The coup de grâce for the nested clade phylogeographic analysis? *Molecular Ecology*, 17:516–518.

Pinho, C. and Hey, J. (2010). Divergence with gene flow: models and data. *Ann. Rev. Ecol. Evol. Syst.*, 41:215–230.

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D., and Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, 55:595–609.

Porter, A. H. (1990). Testing nominal species boundaries using gene flow statistics: The taxonomy of two hybridizing admiral butterflies (Limenitis: Nymphalidae). *Systematic Zoology*, 39:131.

Puillandre, N., Lambert, A., Brouillet, S., and Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular ecology*, 21:1864–1877.

Rannala, B. (2015). The art and science of species delimitation. *Current Zoology*, 61:846–853.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164:1645–1656.

Rannala, B. and Yang, Z. (2013). Improved reversible jump algorithms for Bayesian species delimitation. *Genetics*, 194:245–253.

Rannala, B. and Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, 66:823–842.

Rogers, A. R. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, 9:552–569.

Sites Jr, J. W. and Marshall, J. C. (2004). Operational criteria for delimiting species. *Annu. Rev. Ecol. Evol. Syst.*, 35:199–227.

Smith, M. A., Fisher, B. L., and Hebert, P. D. N. (2005). Dna barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of madagascar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1825–1834.

Smith, M. L. and Carstens, B. C. (2019). Process-based species delimitation leads to identification of more biologically relevant species. *Evolution*, 74(2):216–229.

Sneath, P. H. A. (1957a). The application of computers to taxonomy. *Microbiology*, 17:201–226.

Sneath, P. H. A. (1957b). Some thoughts on bacterial classification. *Journal of General Microbiology*, 17:184–200.

Sneath, P. H. A. (1976). Phenetic taxonomy at the species level and above. *Taxon*, pages 437–450.

Solís-Lemus, C., Knowles, L. L., and Ané, C. (2015). Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution*, 69(2):492–507.

Sukumaran, J. and Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. *Proc. Natl. Acad. Sci. U.S.A.*, 114(7):1607–1612.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., and Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends in Ecology & Evolution*, 18:70–74.

Templeton, A. R. (2001). Using phylogeographic analyses of gene trees to test species status and processes. *Molecular Ecology*, 10:779–791.

Wayne, L., Brenner, D., Colwell, R., Grimont, P., Kandler, O., Krichevsky, M., Moore, L., Moore, W., Murray, R., Stackebrandt, E., et al. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic and Evolutionary Microbiology*, 37:463–464.

Wiens, J. J. and Servedio, M. R. (2000). Species delimitation in systematics: inferring diagnostic differences between species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267:631–636.

Will, K. W. and Rubinoff, D. (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, 20:47–55.

Wilson, K. H. (1995). Molecular biology as a tool for taxonomy. *Clinical Infectious Diseases*, 20:S117–S121.

Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In *Soft computing and industry*, pages 25–42. Springer.

Wu, D.-D., Ding, X.-D., Wang, S., Wojcik, J. M., Zhang, Y., Tokarska, M., Li, Y., Wang, M.-S., Faruque, O., Nielsen, R., Zhang, Q., and Zhang, Y.-P. (2018). Pervasive introgression facilitated domestication and adaptation in the bos species complex. *Nature Ecol. Evol.*, 2(7):1139–1145.

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Curr. Zool.*, 61(5):854–865. http://dx.doi.org/10.1093/czoolo/61.5.854.

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences USA*, 107:9264–9269.

Yang, Z. and Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12):3125–3135.

Zhang, C., Rannala, B., and Yang, Z. (2014). Bayesian species delimitation can be robust to guide tree inference errors. *Syst. Biol.*, 63(6):993–1004.

Zhang, C., Zhang, D.-X., Zhu, T., and Yang, Z. (2011). Evaluation of a Bayesian coalescent method of species delimitation. *Syst. Biol.*, 60:747–761.

Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29:2869–2876.

Chapter 5.6 A Tutorial on the Use of BPP for Species Tree Estimation and Species Delimitation

Tomáš Flouri[1]
Department of Genetics, Evolution and Environment, University College London
London WC1E 6BT, United Kingdom
t.flouris@ucl.ac.uk
https://orcid.org/0000-0002-8474-9507

Bruce Rannala
Department of Evolution and Ecology, University of California Davis
One Shields Avenue, Davis CA USA
brannala@ucdavis.edu
https://orcid.org/0000-0002-8355-9955

Ziheng Yang[2]
Department of Genetics, Evolution and Environment, University College London
London WC1E 6BT, United Kingdom
z.yang@ucl.ac.uk
https://orcid.org/0000-0003-3351-7981

Abstract

BPP is a Bayesian Markov chain Monte Carlo program for analyzing multilocus sequence data under the multispecies coalescent (MSC) model with and without introgression. Among the analyses that can be conducted are estimation of population size and species divergence times, species tree estimation, species delimitation and estimation of cross-species introgression intensity. The program can also be used to simulate gene trees and sequence alignments under the MSC model with, or without, migration. In this tutorial, we illustrate the use of BPP for species tree estimation and species delimitation. We also provide practical guidelines on running BPP on multicore systems. As BPP is continuously updated, the most up-to-date version of this tutorial, as well as the data files, are available at `http://github.com/bpp/tutorial`.

How to cite: Tomáš Flouri, Bruce Rannala, and Ziheng Yang (2020). A Tutorial on the Use of BPP for Species Tree Estimation and Species Delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 5.6, pp. 5.6:1–5.6:16. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Supplement Material `https://github.com/bpp/tutorial`

1 Introduction

BPP is a Bayesian Markov Chain Monte Carlo (MCMC) program for analyzing sequence alignments from multiple loci and multiple species under the multispecies coalescent (MSC) model (Rannala and Yang, 2003; Yang, 2002) with and without introgression (Xu and Yang 2016; Chapter 5.5 [Rannala and Yang 2020]). The program allows four types of analysis,

[1] T.F. is supported by a Biotechnology and Biological Sciences Research Council grant (BB/P006493/1).
[2] Z.Y. is supported by a Biotechnology and Biological Sciences Research Council grant (BB/P006493/1).

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 5.6; pp. 5.6:1–5.6:16
A book completely handled by researchers.
No publisher has been paid.

referred to as A00, A01, A10, and A11, and specified by using two variables in the control file (Yang, 2015; Flouri et al., 2018, table 1). Analysis A00 is a within-model inference and is used to estimate the parameters in the MSC or MSC with introgression (MSci) models, such as the species divergence times (τs), population sizes (θs), and the introgression probability at hybridization/introgression events (φs) (Rannala and Yang, 2003; Burgess and Yang, 2008; Flouri et al., 2020a), when the species tree model is given by the user and fixed. The other three analyses are trans-model inferences, in which the Markov chain moves between different models. Analysis A01 is used for species tree inference when the assignments of sequences to species are provided by the user (Yang and Rannala, 2014; Rannala and Yang, 2017). Analysis A10 conducts species delimitation using a user-specified guide tree (Yang and Rannala, 2010), and A11 implements joint species delimitation and species tree inference or unguided species delimitation (Yang and Rannala, 2014; Rannala and Yang, 2017).

The basic parameters in the MSC model include the species divergence times (τs) and population size parameters $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per site per generation so that θ is the average proportion of sites having different bases between two sequences sampled at random from the population. Both τs and θs are measured by the expected number of mutations per site. Given a species tree with s species, there are $s - 1$ divergence times and at most $2s - 1$ population size parameters (contemporary populations with only one sequence sampled have no θ parameter). The goal of analysis A00 is to estimate those parameters when the species tree is fixed. Analyses A01, A10, and A11 compare different models (for more information on the MSci model and a review of the MSC model, see Xu and Yang 2016; Flouri et al. 2020a; Chapters 5.5 and 5.6 [Rannala and Yang 2020; Flouri et al. 2020b]).

The current version (4.2.0) of BPP supports a variety of mutation/substitution models, such as JC69, K80, HKY, F81, F84, T92, TN93, and GTR for DNA sequence data. For simplicity, this tutorial focuses on the analysis of closely related species and uses the JC69 mutation mode.

BPP is written in C and can be compiled to run on the command line on any of the major operating systems including MacOS, Linux, and Windows. A basic knowledge of the command line will be needed. Here we use the Unix command line, and Windows users need to make adjustments accordingly. If you have not used the command line before, please work through one of the following short tutorials first:

- `http://abacus.gene.ucl.ac.uk/software/CommandLine.Windows.pdf`
- `http://abacus.gene.ucl.ac.uk/software/CommandLine.MACosx.pdf`

In this tutorial we begin by describing how to install and execute the BPP program. This is followed by an explanation of the basic format of three input files: the sequence alignment file, the imap file, and the control file. We go through the important variables in the control file to illustrate the specification of the priors and the settings for the MCMC algorithm. We then illustrate the A01 analysis for species tree estimation. Once the species tree is inferred, we will use Analysis A00 to estimate the divergence times (τs) and population sizes (θs) on the fixed species tree. We will show how to combine the MCMC samples from multiple runs to produce a summary of the posterior. In the second part, we will use Analysis A11 to conduct a joint species tree estimation and species delimitation.

1.1 Installation of BPP

BPP is an open-source software available for download on GitHub at `http://github.com/bpp/bpp`. The latest stable executable files for Windows and MacOS can be downloaded

Table 1 The four types of analyses implemented in BPP

speciesdelimitation	speciestree	
	0	1
0	**A00.** Estimation of parameters under the multispecies coalescent model (Yang, 2002; Rannala and Yang, 2003)	**A01.** Inference of species tree when the assignment and delimitation are given (Rannala and Yang, 2017)
1	**A10.** Species delimitation using a fixed guide tree (Yang and Rannala, 2010; Rannala and Yang, 2013)	**A11.** Joint species delimitation and species-tree inference or unguided species delimitation (Yang and Rannala, 2014)

from http://github.com/bpp/releases. If you want the most recent pre-release version you will need to have a C compiler and the git program installed on your computer. You can obtain the source code and compile it using the following steps:

```
mkdir software
cd software
git clone http://github.com/bpp/bpp
cd bpp/src
make
```

This creates an executable file called `bpp`, which should be copied into a folder that is included in the `PATH` environment variable. This will allow us to execute BPP without having to type the full path to the executable. For example,

```
mkdir ~/mybpp
cp bpp ~/mybpp
export PATH=$PATH:~/mybpp
```

For detailed instructions on compilation and installation please see the GitHub repository.

To run an analysis of the the example dataset, create a folder called `bpptutorial/data/` and copy the example data files in the folder from either http://abacus.gene.ucl.ac.uk/ziheng/data.html or http://abacus.gene.ucl.ac.uk/ziheng/data/HornedLizardsData.tgz. Then create another working folder `A00/r1/`, and run the program there. For example,

```
mkdir -p bpptutorial/data
cd bpptutorial/data
wget http://abacus.gene.ucl.ac.uk/ziheng/data/HornedLizardsData.tgz
tar -xvzf HornedLizardsData.tgz
mkdir -p A00/r1
cd A00/r1
bpp --cfile ../../lizards.bpp.A00.ctl
```

2 Input files

A BPP analysis requires three input files: a control file, a multiple sequence alignment file and a imap file. The control file specifies the type of analysis, sets the parameters of the prior distributions and specifies the details of the MCMC run. The sequence alignment file contains aligned sequences for one or more loci, arranged one after another. The Imap file assigns/maps individuals to populations or species.

5.6:4 BPP tutorial

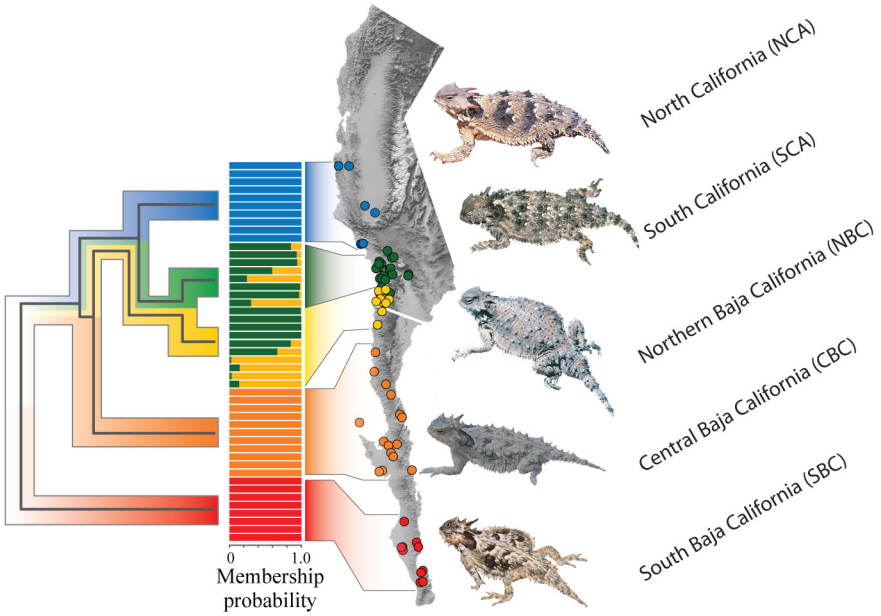


Figure 1 Geographical distributions of Coast Horned Lizards (genus *Phrynosoma*), with five phylogeographic groups arranged latitudinally: North California (NCA), South California (SCA), Northern Baja California (NBC), Central Baja California (CBC), and South Baja California (SBC). Pictures courtesy of Dr Adam Leaché.

2.1 Dataset and multiple sequence alignment file

For this tutorial we will use the Coast Horned Lizard dataset of Leaché et al. (2009), which includes two nuclear loci (*BDNF*: 132 sequences, 529bp; and *RAG-1*: 136 sequences, 1100 bp). This dataset was analyzed by Yang and Rannala (2014) using an earlier version of BPP. Assignment is based on an mtDNA phylogeny, with five phylogeographic groups arranged latitudinally: North California (NCA), South California (SCA), Northern Baja California (NBC), Central Baja California (CBC), and South Baja California (SBC) (Figure 1). Hence, there are five species or populations in the BPP analysis.

The multiple sequence alignment file is a single file that contains the sequence data for all loci in sequential PHYLIP format, with one alignment followed by another. Each sequence alignment (for one locus) is specified using the PHYLIP sequential format, with the first line for each locus specifying the number of sequences and the number of sites in the alignment (see Figure 1, left). Subsequent lines specify the sequences: each line starts with the sequence name followed by two or more whitespaces and then the sequence itself. The sequence name consists of two parts, in the format `x^y` where `x` is a sequence name while `y` is a tag for the specimen or individual. The individual is then assigned to a population or species in the Imap file. A portion of the sequence alignment file and Imap file is shown in Listing 1. Listing 2 depicts the control file we will use for the species tree estimation tutorial.

2.2 Imap file

The Imap file has two columns, separated by white spaces: the first column contains a unique label for each individual and the second column contains the population (or species) name the individual is assigned to. Each individual that occurs in the sequence alignment file must

```
phryno.txt                                          phryno5s.Imap.txt

136 1054                                            1 CBC
BCN_10a^1    ATAAAGGAAAAGCGGCAGCT...                2 CBC
BCN_10b^2    ATAAAGGAAAAGCGGCAGCT...                3 NBC
BCN_11a^3    ATAAAGGAAAAGCGGCAGCT...                4 NBC
BCN_11b^4    ATAAAGGAAAAGCGGCAGCT...                5 NBC
BCN_14a^5    ATAAAGGAAAAGCGGCAGCT...                6 NBC
BCN_14b^6    ATAAAGGAAAAGTGGCAGCT...                7 CBC
...                                                 ...
```

Listing 1 Portions of the sequence data file (`phryno.txt`) and the Imap file (`phryno5s.Imap.txt`) on the left and right, respectively. In the sequence data file each sequence is tagged (1, 2, etc). The part of the sequence name before the caret (^) is read and then ignored. In the Imap file each individual tag is assigned to a population.

be assigned to a population in the Imap file, although the Imap file may include individuals for which no sequence data are available. See Listing 1, right.

3 The control file and BPP settings

See Listing 2 for an example control file. A complete reference of options in the control file is available on the BPP GitHub wiki page at http://github.com/bpp/bpp/wiki. Here, we provide a description of those options relevant to this tutorial. The general format for most options in the control file is: `Option = value(s)`. Text of a line following a symbol '#' or '*' is considered a comment and ignored.

The four types of analysis are specified by setting two binary variables: `speciesdelimitation` and `speciestree`. Those variables take values 0 (meaning disabled) or 1 (enabled). Table 1 illustrates the combination of variables triggering the corresponding analyses.

```
          seed =  -1
       seqfile = ../phryno.txt
      Imapfile = ../phryno5s.Imap.txt
       outfile = out.txt
      mcmcfile = mcmc.txt
* speciesdelimitation = 0 * fixed species tree
* speciesdelimitation = 1 0 2 * delimitation algorithm0 finetune(e)
* speciesdelimitation = 1 1 2 0.5 * delimitation algorithm1 finetune(a m)
         speciestree = 1 * species-tree by SPR

 speciesmodelprior = 1          * 0: uniform labeled histories; 1:uniform rooted trees

  species&tree = 5  NCA SCA NBC CBC SBC
                     18  44  20  34  20
                 ((NCA, ((SCA, NBC), CBC)), SBC);

       usedata = 1    * 0: no data (prior); 1:seq like
         nloci = 2    * number of data sets in seqfile

     cleandata = 0    * remove sites with ambiguity data (1:yes, 0:no)?

    thetaprior = 3 0.004 e * Inv-Gamma(a, b) for theta
      tauprior = 3 0.004   * Inv-Gamma(a, b) for root tau & Dirichlet(a) for other tau's

      * auto (0 or 1): finetune for GBtj, GBspr, theta, tau, mix, locusrate, seqerr
      finetune = 1: .01 .0001 .005 .0005 .2 .01 .01 .01

         print = 1 0 0 0   * MCMC samples, locusrate, heredityscalars Genetrees
        burnin = 8000
      sampfreq = 2
       nsample = 100000
       threads = 2 1 1
```

Listing 2 Sample control file `lizards.bpp.A01.ctl` for species tree estimation (with `speciesdelimitation = 0` and `speciestree = 1`. Lines starting with an asterisk are comments and the default values of `speciesdelimitation` and `speciestree` are 0.

Option `seed` should be a positive integer and sets the seed for the pseudo-random number generator. Runs with identical seeds analyzing the same data will produce identical results. Setting seed to -1 will cause BPP to use a randomly generated seed (which is recorded in the output file `Seedused`). Option `species&tree` defines the species and the starting species tree, and is typically specified in three lines with the following syntax:

```
species&tree = S  S_1 S_2 ... S_S
                  N_1 N_2 ... N_S
                 NEWICK-TREE
```

If only one population/species is specified, the last line (NEWICK-TREE) must be left empty. In the first line we define the number of species S followed by a list of the species names (S_1 to S_S) separated by whitespaces. The second line comprises S numbers, where number N_i indicates the maximum number of sequences for species S_i at any locus (the actual number of sequences at any locus must be less than or equal to this value). Finally, the last line is the Newick (https://en.wikipedia.org/wiki/Newick_format) representation for the starting species tree.

BPP uses a Bayesian model-selection framework to evaluate different models of species phylogenies and species delimitations. Therefore, we specify prior probabilities for the models (species trees) to be compared. The two priors for species trees implemented in BPP are specified using the `speciesmodelprior` keyword (Table 2) with Prior 0 (`speciesmodelprior = 0`) assigning equal probabilities to labeled histories (rooted trees with the internal nodes ordered by age; see Figure 2 and Chapter 5.5 [Rannala and Yang 2020]) and Prior 1 (`speciesmodelprior = 1`) assigning equal probabilities to rooted trees (Yang and Rannala, 2014).

Table 2 Four species tree/species delimitation priors implemented in BPP (using the control variable `speciesmodelprior`)

Prior	Description
0	Assigns equal probabilities to labeled histories (rooted trees with internal nodes ordered by age)
1	Assigns equal probabilities to rooted species trees
2	Assings equal probabilities for the number of delimited species (that is, $1/s$ each for $1, 2, \ldots, s$ delimited species given s populations) and divides up the probability for any specific number of species among the compatible models of species delimitation and species phylogeny in proportion to the number of compatible label histories
3	Same as Prior 2 but instead divides the probability among the compatible models uniformly

Note.— Priors 0 and 1 are used for species tree estimation (analyses A01) and species delimitation on a guide tree (analysis A10), while priors 0–3 are used for joint species delimitation and species tree estimation (analysis A11). This prior has no effect for analysis A00.

For example, there are 15 rooted trees in the case of four species, with 12 unbalanced and 3 balanced trees (Figure 2). Each unbalanced tree, e.g., $(((a, b), c), d)$, is compatible with only one labeled history as there is only one ordering of the internal nodes. Each balanced tree, e.g., $((a, b), (c, d))$, is compatible with two labeled histories, depending on whether the ancestor of a and b is older or younger than the ancestor of c and d. Prior 0 assigns the probability 1/18 to each of the unbalanced trees and 2/18 to each of the balanced trees. Prior 1 assigns the probability 1/15 to each of the 15 rooted trees. Here, we use Prior 1, which is the default.

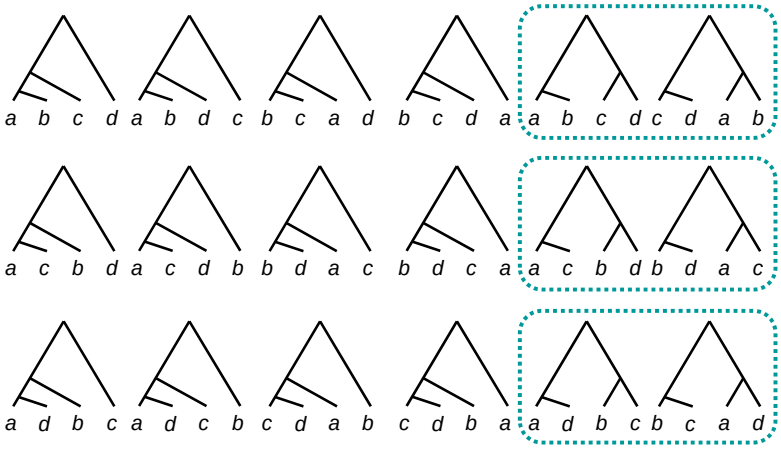


Figure 2 The 18 labeled histories (rooted trees with internal nodes ranked by age) and 15 rooted trees for four tips. Each unbalanced rooted tree is compatible with only one labeled history, but each balanced rooted tree is compatible with two labeled histories.

Within each species tree model, we assign the inverse-gamma priors $\theta \sim \text{IG}(3, 0.004)$ for all θs and $\tau \sim \text{IG}(3, 0.004)$ for the age τ_0 of the root. The inverse-gamma $\text{IG}(a, b)$ has mean $m = b/(a-1)$ if $a > 1$ and variance $s^2 = b^2/[(a-1)^2(a-2)]$ if $a > 2$. If little information is available about the parameters, you can use $a = 3$ for a diffuse prior and then adjust b so that the mean is reasonable. For example, parameter θ measures the genetic diversity (heterozygosity) in the species. This varies among species, with 0.01 (one difference per 100 bps) to be a large value while 0.001 a small value. Parameter τ_0 measures the age of the root in the rooted species tree and depends on the species included in the data set. Thus including an outgroup species will typically mean that a larger prior mean for τ_0 is appropriate.

When specifying the priors, it may be useful to plot the inverse-gamma density and calculate the 95% prior interval. The corresponding R functions are in the MCMCpack, which can be installed with the following command in the R interpreter:

```
install.packages("invgamma");
```

Then we can use the following code in R to plot the inverse-gamma density $\text{IG}(3, 0.004)$ and calculate the 95% prior interval, which is $(0.000554, 0.006465)$:

```
library("invgamma")
a=3; b=0.004;
curve(dinvgamma(x,a,b), from=0, to=0.01)
qinvgamma(c(0.025, 0.975), a, b)
```

Alternatively, a web application may be used for plotting the inverse Gamma (`rdrr.io/cran/bayesAB/man/plotInvGamma.html`). Parameters θ and τ are assigned inverse-gamma priors, using the options `thetaprior` and `tauprior`, respectively. Both options accept two parameters: α and β. In addition, `thetaprior` accepts an optional third parameter - the letter '`e`' (as in estimate). If the third parameter is not specified, BPP integrates out analytically the θ parameters, using conjugate inverse-gamma priors (Hey and Nielsen, 2007). This reduces the state of the Markov chain, resulting in slight improvement in the mixing properties of cross-model MCMC algorithms. The downside is that this approach cannot be used in conjunction with parallelization (see Parallelization).

The number of MCMC iterations is determined by three variables in the control file as: `burnin` + `nsample` × `sampfreq`, where `burnin` is the number of samples that will be discarded (not logged in the sample file) before starting to log a sample every `sampfreq` iterations. The total number of samples logged in the file `mcmc.txt` will be `nsample`.

To assess convergence, run each analysis multiple times (at least twice) using different starting seeds, which are specified in the control file (see Section 2). If the results appear different between runs, re-run the program using a larger number of samples (`nsample`) and/or larger sampling frequency (`sampfreq`). Standard tools are available for diagnosing convergence and mixing problems of MCMC algorithms (pp. 459-510 in Robert and Casella 2005; pp. 226-244 in Yang 2014). However, our experience suggests that running the same analysis multiple times with different seeds and examining the consistency of estimates across runs is the most effective method to guarantee the reliability of the results.

4 Species tree estimation (A01)

We assume the BPP executable resides in a folder which is part of `PATH` (i.e. can be executed without explicitly specifying its path location). We assume that the input files (data, imap and control file) are in a folder named `lizards`. We create two subfolders called `lizards/A01/r1` and `lizards/A01/r2`, and copy the control files to the subfolders,

```
cd lizards
mkdir -p A01/r1 A01/r2
cp lizards.bpp.A01.ctl A01/r1
cp lizards.bpp.A01.ctl A01/r2
```

Note that the `seqfile` (sequence alignment file) and `imapfile` (imap file) variables in the control file (Listing 2) specify that both files are two levels up (e.g., `../../`) in the directory hierarchy relative to the current working directory, while `outfile` and `mcmcfile` specify the current directory (either `A01/r1` or `A01/r2`). Also recall that analysis A01 is triggered by having `speciesdelimitation=0` and `speciestree=1`. We execute each run with the following commands:

```
cd A01/r1
bpp --cfile lizards.bpp.A01.ctl

# Wait until the run is finished

cd ../r2
bpp --cfile lizards.bpp.A01.ctl
cd ../../
```

For the runs we can use the option `threads = 2`. On a laptop computer with a dual-core Intel i7-7500 CPU, one run with two threads takes roughly 10 minutes.

BPP 4 currently uses the subtree pruning and regrafting (SPR) algorithm to search through the space of species tree in the MCMC (Rannala and Yang, 2017). The program collects the sampled species trees (and θs and τs) into the sample file `mcmc.txt`. The summary of the MCMC sample is shown in Listing 3 and the top three species trees (out of a total of 16 in the sample) are illustrated on Figure 3, along with their posterior probabilities. Those three trees have a total posterior probability of 0.96 and therefore consists the 95% credibility set. The majority-rule consensus tree, i.e. the tree built from clades appearing in at least 50% of the trees in the sample, is a binary (resolved) tree and is in line with the maximum a posteriori (map) tree (the tree with highest posterior probability) in the sample.

```
 -3% 0.72 0.31 0.10 0.11 0.41   0.0325   0.0016   0.0014   1467.13645  -2863.66605
Current Pjump:       0.72446   0.31473   0.09694   0.11175   0.40700
Current finetune:    0.01000   0.00010   0.00500   0.00050   0.20000
New finetune:        0.04248   0.00011   0.00151   0.00017   0.29183
 -2% 0.72 0.31 0.25 0.15 0.22   0.0335   0.0015   0.0013   1516.46431  -2860.71283
Current Pjump:       0.72493   0.31210   0.25456   0.14675   0.22400
Current finetune:    0.04248   0.00011   0.00151   0.00017   0.29183
New finetune:        0.18078   0.00011   0.00125   0.00008   0.21028
 -1% 0.73 0.31 0.28 0.28 0.38   0.0220   0.0015   0.0012   1493.60713  -2860.03028

Current Pjump:       0.72533   0.31019   0.27683   0.28025   0.38000
Current finetune:    0.18078   0.00011   0.00125   0.00008   0.21028
New finetune:        0.77068   0.00012   0.00114   0.00007   0.28047
  0% 0.73 0.30 0.29 0.29 0.25   0.0375   0.0015   0.0012   1528.87389  -2859.22283   0:22

Current Pjump:       0.72657   0.29810   0.29267   0.29100   0.25100
Current finetune:    0.77068   0.00012   0.00114   0.00007   0.28047
New finetune:        3.30232   0.00011   0.00111   0.00007   0.22902
  5% 0.72 0.30 0.31 0.28 0.35   0.0313   0.0016   0.0012   1546.63206  -2860.31352   0:53
 10% 0.72 0.30 0.31 0.30 0.35   0.0306   0.0015   0.0012   1489.25672  -2860.20919   1:28
...
 95% 0.72 0.30 0.31 0.29 0.35   0.0358   0.0016   0.0012   1525.98871  -2860.60363   11:25
100% 0.72 0.30 0.31 0.29 0.35   0.0360   0.0016   0.0012   1469.18934  -2860.57826   12:00

12:00 spent in MCMC

Species in order:
   1. NCA
   2. SCA
   3. NBC
   4. CBC
   5. SBC

(A) Best trees in the sample (16 distinct trees in all)
    63037   0.63036   0.63036 ((CBC, ((NBC, SCA), NCA)), SBC);
    19842   0.19842   0.82878 (((CBC, (NBC, SCA)), NCA), SBC);
    12596   0.12596   0.95474 (((CBC, NCA), (NBC, SCA)), SBC);
     2326   0.02326   0.97800 ((CBC, ((NBC, NCA), SCA)), SBC);
     1592   0.01592   0.99392 ((CBC, (NBC, (NCA, SCA))), SBC);
      250   0.00250   0.99642 (((CBC, (NBC, SCA)), SBC), NCA);
      162   0.00162   0.99804 ((CBC, SBC), ((NBC, SCA), NCA));
...

(B) Best splits in the sample of trees (13 splits in all)
 99393 0.993920  11110
 96078 0.960770  01100
 67204 0.672033  11100
 20123 0.201228  01110
 12638 0.126379  10010
  2331 0.023310  10100
  1592 0.015920  11000
...

(C) Majority-rule consensus tree
(((NCA, (SCA, NBC) #0.960770) #0.672033, CBC) #0.993920, SBC);

(D) Best tree (or trees from the mastertree file) with support values
((CBC, ((NBC, SCA) #0.960770, NCA) #0.672033) #0.993920, SBC);   [P = 0.630364]
```

Listing 3 Output from analysis A01 (species tree estimation). The progress indicator is negative during burnin, and BPP goes through four rounds of automatic step-length adjustments, aiming to achieve a near-optimal acceptance proportion of 30% for the parameter moves (Yang and Rodríguez, 2013). Sampling in `mcmc.txt` starts after the burn-in is over. At the end of the MCMC run, the sample is processed to calculate the posterior probabilities of the species trees, which are further summarized to calculate the posterior for splits as well as the majority-rule consensus tree.

In the next step we will run the A00 analysis with the species tree fixed at the MAP tree to estimate the parameters of the MSC model. Using the same control file but with `speciestree=0`, we conduct two runs of the A00 analysis:

```
mkdir -p A00/r1 A00/r2
cp lizards.bpp.A00.ctl A00/r1
cp lizards.bpp.A00.ctl A00/r2
cd A00/r1
bpp --cfile lizards.bpp.A00.ctl

# Wait until the run is finished

cd ../r2
bpp --cfile lizards.bpp.A00.ctl
cd ../
```

We can combine the output of the two runs to increase the number of samples from the posterior distribution and summarize the new concatenated sample independently. To do that we create a new folder and copy the MCMC sample file of the first run, and then concatenate the samples from the second run (note the `tail -n +2` command which skips the header line from the mcmc sample file):

```
mkdir combined; cd combined
cp ../r1/lizards.bpp.A00.ctl .
cp ../r1/mcmc.txt .
tail -n +2 ../r2/mcmc.txt >> mcmc.txt

# !IMPORTANT! Edit print line in control file to read print=-1
bpp --cfile lizards.bpp.A00.ctl
```

Lastly, we must change the line `print = 1 0 0 0 0` in the control file to `print = -1`. This causes BPP to only read and summarize the specified MCMC sample file rather than running a new MCMC analysis. We again run BPP and the posterior means are shown in Figure 4 along with the 95% credible interval for divergence times for each internal node.

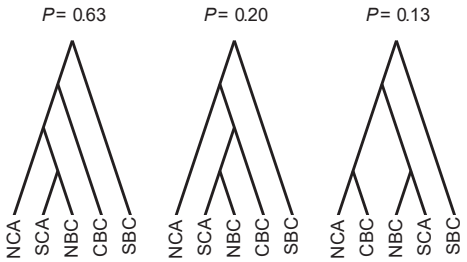


Figure 3 The top three species trees in the 95% CI and their posterior probabilities, with a total probability of 0.96

5 Species delimitation (A11)

In Analysis A11, both the species delimitation model and the species phylogeny are changing in the MCMC. We change the variables in the control file to have `speciesdelimitation = 1` and `speciestree = 1`, create the necessary directory structure and re-run BPP in the

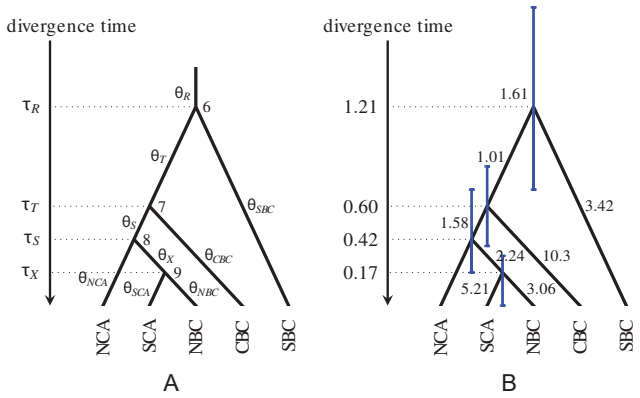


Figure 4 A species tree for five horned lizard species/populations NCA, SCA, NBC, CBC and SBC, illustrating the parameters in the multispecies coalescent model. Those include four species divergence time parameters (τ) for the three ancestral nodes, 6 (NCA, SCA, NBC, CBC, SBC), 7 (NCA, SCA, NBC, CBC), 8 (NCA, SCA, NBC) and 9 (SCA, NBC), and nine population size parameters (θs) for the nine populations on the tree. Estimates were multiplied by 10^3.

same way as in the A01 analysis. The control file differs from the A01 file only in the value of the `speciesdelimitation` variable.

```
cd lizards
mkdir -p A11/r1 A11/r2
cp lizards.bpp.A11.ctl A11/r1
cp lizards.bpp.A11.ctl A11/r2

cd A11/r1
bpp --cfile lizards.bpp.A11.ctl
cd ../r2
bpp --cfile lizards.bpp.A11.ctl
```

The algorithm explores different species delimitation models and different species phylogenies. The assignment to populations is nevertheless fixed; that is, the program attempts to merge different populations into one species but never tries to split one population into multiple species. The SPR algorithm is used to change the species tree topology (Rannala and Yang, 2017), while a reversible-jump MCMC (rjMCMC) algorithm is used to move between different species delimitation models, by either splitting one species into two or joining two species into one species (Yang and Rannala, 2010). Two alternative rjMCMC algorithms are implemented in BPP, which differ in the way that new θ parameters are proposed during the split move. They are specified through the `speciesdelimitation` option which takes one of two formats:

```
speciesdelimitation = 1 0 ε # Algorithm 0
speciesdelimitation = 1 1 a m # Algorithm 1
```

The second digit (0 or 1) distinguishes between the two rjMCMC algorithms. For Algorithm 0, we use a value of $\varepsilon = 2$ in equations 3 and 4 of Yang and Rannala (2010). Reasonable values for ε are 1, 2, 5, etc. For Algorithm 1, we set $a = 2$ and $m = 1$ in equations 6 and 7 of Yang and Rannala (2010). Reasonable values are $a = 1, 1.5, 2$ etc., and $m = 0.5, 1, 2$ etc. When the chain mixes well, the results should be the same between multiple runs and between the two algorithms.

5.6:12 BPP tutorial

BPP offers four priors on delimitation models for Analysis A11, specified using the variable `speciesmodelprior`, which takes the values 0, 1, 2, or 3, with Prior 1 being the default. These are outlined in Table 2. Prior 3 may be suitable when there is a large number of populations. One such scenario is when each sequence (specimen) is assigned into its own "population", so that BPP will explore different models of assignment, species delimitation and species tree estimation (Olave et al., 2014). For this tutorial we use the default Prior 1 (for more information on Priors 2 and 3 see Yang, 2015).

The runs should take 10 to 15 minutes on a modern laptop when using 2 threads. The program collects the sampled species trees (and θs and τs) into the sample file `mcmc.txt`, as well as the number of delimited species. The branch lengths of the species tree are used to distinguish collapsed populations. Once the analyses finish, we check that the runs have converged by comparing the list of best models and the posteriors on the number of species. The summary of one of the runs is shown on Listing 4. The posterior probability of 5 species (NCA, SCA, NBC, CBC, SBC) is 0.93, while that of four species (with NBC and SCA joined into one species) is 0.07. The results regarding the phylogenetic relationships among the delimited species are consistent with the results of the A01 method, with the 99% credibility set including four distinct species tree topologies, with a posterior probability of 0.58 for the best tree. The data seem to contain more information about species delimitation than about species phylogeny.

```
 -3% 0.73 0.29 0.10 0.04 0.46    4 10 0.0010 0.0000 \
 P(3)=0.4745 0.0015   0.0016   1304.39469 -2863.21212
  Current Pjump:      0.72606   0.29288   0.09978   0.04008   0.45500
  Current finetune:   0.01000   0.00010   0.00500   0.00050   0.20000
  New finetune:       0.04276   0.00010   0.00155   0.00006   0.34061
 -2% 0.70 0.30 0.26 0.43 0.20    3  7 0.0015 0.0000 \
 P(3)=0.6525 0.0014   0.0013   1270.04377 -2857.17969
  Current Pjump:      0.70270   0.30469   0.25726   0.42525   0.19600
  Current finetune:   0.04276   0.00010   0.00155   0.00006   0.34061
  New finetune:       0.16644   0.00010   0.00130   0.00010   0.21257
 -1% 0.71 0.31 0.27 0.29 0.40    5 13 0.0028 0.0095 \
 P(5)=0.5740 0.0015   0.0013   1505.49123 -2859.64272
  Current Pjump:      0.71257   0.30731   0.27124   0.28854   0.39800
  Current finetune:   0.16644   0.00010   0.00130   0.00010   0.21257
  New finetune:       0.67367   0.00010   0.00116   0.00009   0.30111
 0% 0.73 0.31 0.30 0.21 0.24    5 13 0.0000 0.0480 \
 P(5)=1.0000 0.0018   0.0011   1490.58826 -2860.30822   0:26
  Current Pjump:      0.72664   0.31391   0.30472   0.20950   0.24250
  Current finetune:   0.67367   0.00010   0.00116   0.00009   0.30111
  New finetune:       2.88751   0.00011   0.00118   0.00006   0.23667
 5% 0.72 0.30 0.30 0.33 0.34    5 13 0.0000 0.0347 \
 P(5)=1.0000 0.0018   0.0012   1472.90507 -2860.39423   1:01
 10% 0.72 0.31 0.30 0.34 0.34    5 13 0.0000 0.0379 \
 P(5)=1.0000 0.0017   0.0012   1507.73604 -2860.41155   1:36
  ...
 95% 0.72 0.31 0.29 0.34 0.33    5 13 0.0007 0.0345 \
 P(5)=0.9268 0.0016   0.0012   1542.82446 -2860.30955   11:51
 100% 0.72 0.31 0.29 0.33 0.33    5 13 0.0007 0.0353 \
 P(5)=0.9305 0.0016   0.0012   1505.18890 -2860.32580   12:27

 12:27 spent in MCMC

(A) List of best models (count postP #species SpeciesTree)
 58409 0.584090 0.584090 5  (CBC NBC NCA SBC SCA)  ((CBC, ((NBC, SCA), NCA)), SBC);
 19842 0.198420 0.782510 5  (CBC NBC NCA SBC SCA)  (((CBC, (NBC, SCA)), NCA), SBC);
 11567 0.115670 0.898180 5  (CBC NBC NCA SBC SCA)  (((CBC, NCA), (NBC, SCA)), SBC);
 5459 0.054590 0.952770 4  (CBC NBCSCA NCA SBC)  ((CBC, (NBCSCA, NCA)), SBC);
 1383 0.013830 0.966600 5  (CBC NBC NCA SBC SCA)  ((CBC, ((NBC, NCA), SCA)), SBC);
 312 0.013120 0.979720 5  (CBC NBC NCA SBC SCA)  ((CBC, (NBC, (NCA, SCA))), SBC);
 760 0.007600 0.987320 4  (CBC NBCSCA NCA SBC)  (((CBC, NCA), NBCSCA), SBC);
 729 0.007290 0.994610 4  (CBC NBCSCA NCA SBC)  (((CBC, NBCSCA), NCA), SBC);
  ...
(B) 2 species delimitations & their posterior probabilities
 93048 0.930480    5 (CBC NBC NCA SBC SCA)
 6952 0.069520    4 (CBC NBCSCA NCA SBC)

(C) 6 delimited species & their posterior probabilities
 100000 1.000000 SBC
 100000 1.000000 NCA
 100000 1.000000 CBC
 93048 0.930480 SCA
 93048 0.930480 NBC
 6952 0.069520 NBCSCA

(D) Posterior probability for # of species
 P[1] = 0.000000   prior[1] = 0.175000
 P[2] = 0.000000   prior[2] = 0.175000
 P[3] = 0.000000   prior[3] = 0.225000
 P[4] = 0.069520   prior[4] = 0.250000
 P[5] = 0.930480   prior[5] = 0.175000
```

Listing 4 Output from BPP analysis A11 (joint species delimitation and species-tree estimation).

We note that two approaches to species delimitation are implemented in BPP. The first is the approach of Bayesian model comparison, as illustrated above in the A11 analysis (Yang and Rannala, 2010; Leaché et al., 2019). The second is to use the estimates of parameters in the MSC or MSci models and rely on heuristic criteria such as the genealogical divergence index (*gdi*) of Jackson et al. (2017). This approaches relies on the A00 analysis to estimate parameters – given the parameter values, calculation of heuristic index is straightforward. Leaché et al. (2019) suggested a recursive procedure to apply *gdi* when the data include more than two populations. See Chapter 5.5 (Rannala and Yang 2020) for more details.

6 Parallelization

BPP implements two levels of parallelization at the moment: instruction-level and intra-node (or multithreading) parallelism.

6.1 Instruction-level parallelism

Instruction-level parallelism (also known as *vectorization*) is achieved through single-instruction, multiple-data (SIMD) instruction set extensions to the x86 architecture. Currently BPP utilises code for three such instruction sets: *S*treaming SIMD Extensions (SSE), *A*dvanced Vector eXtensions (AVX) and AVX-2. SIMD instructions make use of *vector registers* (storage space within the processor) with a length of 128 (SSE), 256 (AVX and AVX-2) and 512 (AVX-512; not yet supported by BPP) bits. Those registers can hold multiple, independent data values of smaller size (e.g., a 256-bit register can hold four 64-bit double-precision floating-point values). For instance, if two registers contain four values each, the software can perform element-wise multiplications of the vector elements using a single instruction, instead of performing four separate multiplications as in the traditional x86 instruction set. Those instruction sets can significantly speed-up computation in matrix manipulations.

BPP automatically detects the best instruction set available on the computer it is executed on, and uses optimized code for that particular instruction set. On modern hardware, auto-detection works well. However, one can force a specific instruction set using the `arch` option in the control file, which takes four possible values: `CPU`, `SSE`, `AVX`, and `AVX2`. For example, to disable vectorization completely add the following line to the control file:

```
arch = CPU
```

6.2 Multithreading and NUMA architecture

BPP implements multithreading via pthreads and currently supports *medium-grained* parallelization across loci. Loci are distributed evenly to threads, and each thread handles its assigned loci sequentially for each move in an MCMC iteration. A move here is a collection of MCMC proposals of similar nature: for instance, the gene tree node age move cycles through all nodes on the gene tree for a locus and proposes a change to the age of each node. Communication between threads is reduced to one synchronization barrier at the end of each move. The number of threads used cannot exceed the number of loci in the dataset.

Furthermore, modern multiprocessor computers typically utilize the non-uniform memory access (NUMA) architecture, in which memory access time depends on the memory location relative to the processor. Figure 5 depicts the memory layout of a NUMA multiprocessor system with four CPUs. Each CPU may comprise several cores, and has its own local memory which is faster to access than the local memory of another processor. Currently BPP does

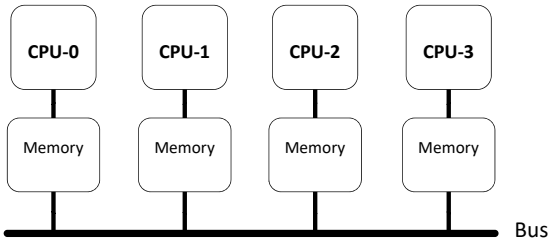


Figure 5 Non-uniform memory access (NUMA) is standard in modern multiprocessing computer architecture, in which memory access is faster if the memory is local to the processor.

not take full advantage of the NUMA memory layout, as all memory accessed throughout the run of the program is allocated locally to the processor on which the first (master) thread is running. Therefore, to achieve optimal performance it is important to ensure that all threads are allocated on cores of the same processor. Given that the typical strategy followed by operating systems is to distribute the processing workload equally across processors, the performance of BPP can degrade substantially when increasing numbers of processors (not to be confused with cores) are involved in the computation. To alleviate this issue, we have implemented core pinning, i.e. each thread is pinned to a particular CPU core.

Multithreading is enabled by specifying the `threads` variable in the control file which has the format `threads = N A B`, where N is the number of threads to be used, A is the starting core/thread number, and B is the stride, so the N threads will be assigned to cores $A, A+B, \ldots, A+(N-1)B$. Parameters A and B are optional, and their default values are 1.

```
threads = 4       * equivalent to threads = 4 1 1
```

The `lscpu` program is available on most GNU/Linux and MacOS distributions and can be used to see the topology of the system, such as the number of processors, cores and threads. For example, the following shows the output of the `lscpu` command for a quad-processor Lenovo ThinkSystem SR850 with four Intel Xeon Gold 6154 CPUs.

```
lscpu | egrep 'NUMA|Thread|Core'
```

and observe the output:

```
Thread(s) per core:    2
Core(s) per socket:    18
NUMA node(s):          4
NUMA node0 CPU(s):     0-17,72-89
NUMA node1 CPU(s):     18-35,90-107
NUMA node2 CPU(s):     36-53,108-125
NUMA node3 CPU(s):     54-71,126-143
```

There are four processors (NUMA nodes), and each processor comprises 18 physical cores, each of which can execute two threads (hyperthreading). Cores 1-18 are part of CPU1, 19-36 of CPU2, 37-54 of CPU3 and 55-72 of CPU4. The remaining 72 cores are hyperthreaded. (Note that the `lscpu` output starts from 0 while we start from 1 in the `threads` option in BPP.) Thus the following uses all 18 cores of the second processor:

```
threads = 18 19
```

or equivalently

```
threads = 18 19 1
```

A second example is for a dual-processor Dell PowerEdge T640 with two Intel Xeon Gold 5118 CPUs.

```
lscpu | egrep 'NUMA|Thread|Core'

Thread(s) per core:  2
Core(s) per socket:  12
NUMA node(s):        2
NUMA node0 CPU(s):   0,2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32,34,36,38,40,42,44,46
NUMA node1 CPU(s):   1,3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33,35,37,39,41,43,45,47
```

In this case, the cores of a processor are not enumerated sequentially, but are interleaved. Then `threads = 4 1 2` will specify the first four cores of CPU 1. The option `threads = 4 1 1` would use the first two cores of CPU 1 and the first two cores of CPU 2, and should be avoided.

Note that using more cores or threads, although always taking more computing resources, may not always reduce the running time. For many datasets involving sequence data from closely related species, we found that using 4 and 8 threads on the same processor gave near optimal performance. A good strategy is to execute a short run with low numbers for `burnin`, `sampfreq` and `nsample` (so that the run finishes in a few minutes) and experiment with different numbers of threads (with `threads` = 1, 2, 4, or 8, say), recording the running time to determine the optimal choice.

7 Discussion

This chapter has outlined the basic features of the BPP program and provided examples of simple analyses aimed at either species tree inference or species delimitation. Our goal has been to provide practical instruction on the use of the program. More detailed information regarding the underlying models implemented in BPP may be found in Chapters 3.3 and 5.5 (Rannala et al. 2020; Rannala and Yang 2020).

Detailed BPP documentation describing all the features and options of the program is available on the GitHub wiki at `https://github.com/bpp/bpp/wiki` and as a PDF manual. User support is available on the BPP Google group at `https://groups.google.com/forum/#!forum/bpp-discussion-group`. A web application for preparing BPP control and map files is available at `https://brannala.github.io/bpps/`.

References

Burgess, R. and Yang, Z. (2008). Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.*, 25:1979–1994.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2018). Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, 35(10):2585–2593.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2020a). A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.*, 37(4):1211–1223.

Flouri, T., Rannala, B., and Yang, Z. (2020b). A tutorial on the use of bpp for species tree estimation and species delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.6, pages 5.6:1–5.6:16. No commercial publisher | Authors open access book.

Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. U.S.A.*, 104(8):2785–2790.

Jackson, N. D., Carstens, B. C., Morales, A. E., and O'Meara, B. C. (2017). Species delimitation with gene flow. *Syst. Biol.*, 66(5):799–812.

Leaché, A. D., Koo, M. S., Spencer, C. L., Papenfuss, T. J., Fisher, R. N., and McGuire, J. A. (2009). Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (*Phrynosoma*). *Proc. Natl. Acad. Sci. U.S.A.*, 106:12418–12423.

Leaché, A. D., Zhu, T., Rannala, B., and Yang, Z. (2019). The spectre of too many species. *Syst. Biol.*, 68(1):168–181.

Olave, M., Solà, E., and Knowles, L. L. (2014). Upstream analyses create problems with DNA-based species delimitation. *Systematic Biology*, 63(2):263–271.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164:1645–1656.

Rannala, B. and Yang, Z. (2013). Improved reversible jump algorithms for Bayesian species delimitation. *Genetics*, 194(1):245–253.

Rannala, B. and Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, 66:823–842.

Rannala, B. and Yang, Z. (2020). Species delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.5, pages 5.5:1–5.5:18. No commercial publisher | Authors open access book.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer-Verlag, Berlin, Heidelberg.

Xu, B. and Yang, Z. (2016). Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, 204:1353–1368. doi: 10.1534/genetics.116.190173.

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4):1811–1823.

Yang, Z. (2014). *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford, England.

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Curr. Zool.*, 61(5):854–865. http://dx.doi.org/10.1093/czoolo/61.5.854.

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. U.S.A.*, 107:9264–9269.

Yang, Z. and Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12):3125–3135.

Yang, Z. and Rodríguez, C. E. (2013). Searching for efficient Markov chain Monte Carlo proposal kernels. *Proc. Natl. Acad. Sci. U.S.A.*, 110(48):19307–19312.