OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes
Celine Scornavacca, Khalid Belkhir, Jimmy Lopez, Rémy Dernat, Frédéric Delsuc, Emmanuel Douzery, Vincent Ranwez

To cite this version:
Celine Scornavacca, Khalid Belkhir, Jimmy Lopez, Rémy Dernat, Frédéric Delsuc, et al.. OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. Molecular Biology and Evolution, Oxford University Press (OUP), 2019, 36 (4), pp.861-862. 10.1093/molbev/msz015. hal-02139915

HAL Id: hal-02139915
https://hal.archives-ouvertes.fr/hal-02139915
Submitted on 26 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes

Celine Scornavacca,*1 Khalid Belkhir,1 Jimmy Lopez,1 Rémy Dernat,1 Frédéric Delsuc,1 Emmanuel J.P. Douzery,1 and Vincent Ranwez2

1Institut des Sciences de l’Evolution (ISEM), Univ. Montpellier, CNRS, EPHE, IRD, Montpellier, France
2AGAP, Univ. Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

*Corresponding author: E-mail: celine.scornavacca@umontpellier.fr.
Associate editor: Koichiro Tamura

Abstract

We present version 10 of OrthoMaM, a database of orthologous mammalian markers. OrthoMaM is already 11 years old and since the outset it has kept on improving, providing alignments and phylogenetic trees of high-quality computed with state-of-the-art methods on up-to-date data. The main contribution of this version is the increase in the number of taxa: 116 mammalian genomes for 14,509 one-to-one orthologous genes. This has been made possible by the combination of genomic data deposited in Ensembl complemented by additional good-quality genomes only available in NCBI. Version 10 users will benefit from pipeline improvements and a completely redesigned web-interface.

Key words: orthologous sequences, mammals, coding sequences, phylogenomics, comparative genomics.

OrthoMaM is a database of high-quality orthologous sequence alignments and phylogenetic trees from mammalian genomes. It has been, for instance, used for developing new molecular markers, inferring mammalian phylogenies, simulating sequences for testing alignment filtering methods, and studying the evolution of base composition in protein-coding sequences. Previous versions of our database only included mammalian genomes from Ensembl (Ranwez et al. 2007; Douzery et al. 2014). With the progress of sequencing techniques, the number of genomes included in NCBI, but not yet in Ensembl, grows each year. This motivated us to totally rethink our database to include the annotated mammalian genomes available in NCBI only.

To identify the core set of orthologous sequences to be used, we rely on the OrthoMaM v8 pipeline described in Douzery et al. (2014). Briefly, we isolate 1-to-1 orthologous genes among pairs of our pillar placental species (Homo-Mus, Homo-Canis, and Mus-Canis) by using Ensembl v91 annotations (Zerbino et al. 2018). These clusters are then enriched by adding sequences of 69 additional mammals annotated in Ensembl v91 as 1-to-1 orthologues to the human gene, and turned into clusters of 1-to-1 orthologous CDSs by selecting the longest transcript of each gene. These proto-clusters are aligned using MAFFT at the amino acid level (Kato and Standley 2013).

The CDSs of the 47 annotated mammal genomes available in NCBI (Rigden and Fernández 2017) but not in Ensembl as of March 2018 are then used to enrich the proto-alignments. For each proto-alignment, an HMM profile is created via hmmsearch using default parameters of the HMMER toolkit (Eddy 2011). Additionally, all HMM profiles are concatenated and summarized using hmmpress to construct an HMM database. Then, for each NCBI CDS Cj, hmmscan is used on the HMM database to get the best hits among the proto-alignments—denoted bestAl(Cj). For each proto-alignment Aij, the most similar sequences for each species S—denoted bestSeq(S, Aij)—are detected via hmmssearch. Outputs from hmmssearch and hmmscan are discarded if the first hit score is not substantially better than the second (hit2 < 0.9 hit1) and are combined in a best-reciprocal-hit fashion: Suppose that we are given a NCBI CDS sequence Cj, for species Sj; then denoting Aij the best proto-alignment for Cj if the first sequence of bestSeq(Sj, bestAl(Cj)) equals Cj and none of the other sequences Ck ≠ Cj in bestSeq(Aij) is such that bestAl(Sk, Cj) = Aij, then Cj is considered to belong to Aij. This ensures our orthology predictions for the NCBI CDSs to be robust. The enriched orthologous clusters have been thoughtfully aligned and filtered via the OMML_MACSE pipeline to construct high-quality codon alignments relying on MACSE v2 (Ranwez et al. 2018), MAFFT (Katoh and Standley 2013), and HMMcleaner (Philippe et al. 2017). Phylogenetic trees for each CDS alignment are constructed under maximum likelihood (ML) with RAxML (Stamatakis 2014) under the GTR+Γ model (Yang 1994).

OrthoMaM v10 also includes an ad hoc nonorthologous sequence detection method. For a given marker, the distance between the most recent common ancestor (MRCA) of placental sequences and any sequence of the marker, denoted here S0, depends mostly on two factors: the considered marker (some genes evolve faster) and the genome to which S0 belongs (some species evolve faster). We use linear regression to explain the patristic distance on inferred ML trees between a sequence.

© The Author(s) 2019. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com


Open Access
Si and the placental MRCA based on those two factors. When the linear regression prediction departs from the observed patristic distance (standardized residual $>3$), we consider the sequence as spurious and remove it.

As in the previous versions of our database, we use exon positions in CDS alignments to infer exon orthology and alignments (see Douzery et al. 2014 for details). Phylogenetic trees for each exon alignment are reconstructed as for CDSs.

OrthoMaM v10 database is available at http://www.orthomam.univ-montp2.fr/orthomam_v10/. For each CDS and exon marker, we provide gene level information (gene name, GO annotation), full sequence traceability information (sequence identifier in Ensembl/NCBI, filtering details), nucleotide and amino acid alignments, phylogenetic trees, as well as several evolutionary indicators such as relative evolutionary rate and $G+C$ content. The improved web-interface (see fig. 1) provides a user-friendly way to query the database based on alignment content (number of sequences, number of species that should be represented, and alignment length), evolutionary indicators, or sequence similarity using blast.

This new reactive web site, which has been completely redesigned to improve the user experience, the triplication of the number of species, the major update of the pipeline for orthology prediction and sequence filtering should make OrthoMaM v10 a central resource for anyone interested in mammalian comparative genomics.

**Acknowledgments**

Our project largely benefited from the Montpellier Bioinformatics Biodiversity cluster computing platform. This work was supported by the Agence Nationale de la Recherche “Investissements d’avenir/Bioinformatique” (ANR-10-BINF-01-02 project Ancestrome), Labex CeMEB “Centre Méditerranéen de l’Environnement et de la Biodiversité” (ANR-10-LABX-0004), and the European Research Council (ERC-2015-CoG-683257 project ConvergeAnt). This is contribution ISEM 2018-292 of the Institut des Sciences de l’Evolution de Montpellier.

**References**


