



HAL
open science

Towards an integrated ecosystem of R packages for the analysis of population genetic data

Emmanuel Paradis, Thierry Gosselin, Niklaus J. Grünwald, Thibaut Jombart, Stéphanie Manel, Hilmar Lapp

► **To cite this version:**

Emmanuel Paradis, Thierry Gosselin, Niklaus J. Grünwald, Thibaut Jombart, Stéphanie Manel, et al.. Towards an integrated ecosystem of R packages for the analysis of population genetic data. *Molecular Ecology Resources*, 2017, 17 (1), pp.1-4. 10.1111/1755-0998.12636 . hal-01605538

HAL Id: hal-01605538

<https://hal.science/hal-01605538>

Submitted on 23 Jun 2018

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.

November 7, 2016

SPECIAL ISSUE: POPULATION GENOMICS WITH R

Towards an integrated ecosystem of R packages for the analysis of population genetic data

EMMANUEL PARADIS,* THIERRY GOSSELIN,† NIKLAUS J. GRÜNWARD,‡§
THIBAUT JOMBART,¶ STÉPHANIE MANEL** and HILMAR LAPP††

**Institut des Sciences de l'Évolution, Université Montpellier - CNRS - IRD - EPHE, Place Eugène Bataillon – CC 065, 34095 Montpellier cédex 05, France, †Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, QC G1V 0A6, Canada, ‡Horticultural Crops Research Unit, USDA-ARS, Corvallis, OR 97330, USA, §Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA, ¶MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College, London W2 1PG, United Kingdom, **EPHE, PSL Research University, CNRS, UM, SupAgro, IRD, INRA, UMR 5175, CEFE F-34293, Montpellier, France, ††Center for Genomic and Computational Biology (GCB), Duke University, 101 Science Drive, Durham NC 27708, USA*

Email: Emmanuel.Paradis@ird.fr

Word count: 2680 (including title, keywords, references)

Keywords: allelic data, NGS data, R, variant call format

Introduction

As a scientific field, population genetics, despite its relatively young age, occupies a central place in biology. It was considered early on, well before the foundation of molecular genetics, as the way forward to solve the forces behind the evolution of species (Fisher 1930). Today, many pressing issues in understanding or predicting biological responses are investigated through population genetics approaches, for example the adaptation of populations of animals or plants facing human-mediated global changes such as habitat destruction or climate warming (Fordham *et al.* 2014; Merilä & Hendry 2014). Addressing these and other applied questions benefits greatly from an integrative approach that combines analyses of genetic data with geographical and/or ecological data.

The ongoing revolution in sequencing technologies has had a dramatic impact on the questions population geneticists and molecular ecologists can tackle. Genotypic data are now readily available for many loci for many individuals from many microbe, plant, or animal species (Luikart *et al.* 2003; Ellegren 2014). However, acquiring massive amounts of data without an interoperating ecosystem of tools to manipulate, explore, and analyse them properly for a wide variety of research questions results in – paraphrasing John Naisbitt (1984) – “drowning under data and starving from knowledge”. Much progress in addressing this challenge for population genetics has arguably been facilitated through the de facto convergence on a single computer language: the open-source statistical analysis and programming platform R (Ihaka & Gentleman 1996). The main advantages of R for users as well as for developers of analytical methods in population genetics are described elsewhere in this special issue (Kamvar *et al.* 2017; Paradis *et al.* 2017).

Evolutionary ecologists have been among the earliest adopters of R for analysing molecular data: APE (Paradis *et al.* 2004) was first released on CRAN in August 2002, and ADE4 (Dray *et al.* 2007) followed in December of the same year. These and other early adopters, such as ADEGENET, first released in April 2007 (Jombart 2008), created the beginning of an ecosystem that made it increasingly attractive for population geneticists to continue enriching it with packages implementing new methods.

Recognizing the importance of nurturing this ecosystem, and that it faces challenges

from its rapid yet mostly organic growth, the National Evolutionary Synthesis Center (NESCent) held the Population Genetics in R Hackathon on March 16–20, 2015, in
33 Durham (USA) at the Center’s headquarters. Hackathons are intensive coding events that bring together groups of people who normally do not meet, to work collaboratively and face-to-face on a shared objective (Lapp *et al.* 2007; Groen & Calderhead 2015). For
36 the event held at NESCent, the objective was to address interoperability, scalability, and workflow building challenges for users and developers of R packages for population genetics. Such work usually receives little reward in the academic incentive system, and
39 thus tends to fall behind without dedicated initiatives (Howison & Herbsleb 2013; Prins *et al.* 2015). In order to promote the corresponding efforts of not only the participants of the event but also the community as a whole, and to raise awareness of the importance
42 of a high-quality, interoperable, and well-documented ecosystem of analysis tools, this special issue “Population Genomics in R” highlights several products of the hackathon in combination with similar recent works from others in the population genetics community.

45 Hackathons have become increasingly popular, including in scientific computing, to facilitate a range of objectives such as resource adoption, community building, or tool innovation (Trainer *et al.* 2014). Of note, the NESCent hackathon was among those
48 chosen by a Carnegie Mellon University-based research group for studying how and with what success such events use different mechanisms to balance their different objectives. Their results are outside the scope of this special issue and have been reported
51 separately (Trainer *et al.* 2016).

Summary of special issue “Population Genomics in R”

Most papers in this special issue present new packages or significant improvements over
54 existing ones. Two papers deviate from this theme. Kamvar *et al.* (2017) present how new internet-based development platforms enable the community to collaborate on promoting teaching of and education in population genetics methods. Paradis *et al.*
57 (2017) synthesize the recent progress in analysing genomic data for population genetics, and present an overview of the available packages and how they integrate into a common programming environment.

60 A motivation shared by several packages presented in this issue is to provide tools
that better facilitate users' work, particularly handling their data, which may involve
complex sampling designs. STRATAG provides user-friendly tools to handle allelic data
63 from stratified populations, including the possibility to read a variety of data file formats
and perform a series of different analyses (Archer *et al.* 2017). APEX facilitates the
manipulation of sequence data from multiple genes with tools to display them and
66 explore incongruence among them (Jombart *et al.* 2017). GENEPOEDIT is a collection of
tools which helps in the manipulation of large multilocus molecular data sets and
integrates with a variety of other packages (Stanley *et al.* 2017).

69 Another difficult task users often face, especially with multilocus data sets, is data
visualization, which is one of the chief objectives of two packages presented in this issue.
MINOTAUR implements several measures of outliers calculated from high-dimensional
72 genomic data and their visualization with a user-friendly interface (Verity *et al.* 2017).
POPHELPER, which is both an R package and a web server, provides tools for the
visualization of population structure, including the outputs of external applications
75 (Francis 2017).

Handling or analysing (very) big data sets in a scalable manner is increasingly a
challenge with the exponential growth in available data. Addressing this challenge is a
78 common goal among several papers. Knaus & Grünwald (2017) present `vcfr`, a package
for handling variant call format (VCF) files (Danecek *et al.* 2011), including tools to
read, write and visualize VCF as well as FASTA (DNA sequences) and GFF
81 (annotation) files. Paradis *et al.* (2017) present tools from the `PEGAS` package (Paradis
2010) to scan VCF files, select the loci to read, and analyse them using basic R
operations. To handle large quantities of data efficiently, both packages use optimized
84 C/C++ code interfaced with R. Wringe *et al.* (2017) use parallel code execution to
efficiently detect hybrids from multilocus data in their package `PARALLELNEWHYBRID`.

The many different file formats used by population genetics software has been an
87 ongoing interoperability difficulty (discussed in, e.g., Lischer & Excoffier 2012). Many of
the packages presented in this issue have the ability to read many different file formats
(Archer *et al.* 2017; Francis 2017; Paradis *et al.* 2017). However, the general trend
90 among R packages, and other software as well, has been to address the data exchange

format issue by adopting unified file formats: FASTA for DNA sequences and VCF for genotypic data.

93 Testing for or quantifying natural selection in populations is one of the main
objective of population genetics. Two of the papers in this special issue present packages
that test for selection from genomic data, and use innovative algorithmic and
96 implementation approaches to achieve substantial performance improvements over
previous releases. PCADAPT (Luu *et al.* 2017) uses new algorithms for multivariate
analysis in high-dimensional tables (see also Paradis *et al.* 2017), whereas REHH (Vitalis
99 *et al.* 2017) uses multi-threading (light-weight parallelism) for the analysis of SNP data.

Data simulation under more or less complex scenarios has become an increasingly
important task in population genetics. Two papers in this special issue present packages
102 designed for making this task easier. SKELESIM simulates genetic data with a
user-friendly interface to help users to set parameters or choose sample sizes, as well as
tools to summarize outputs (Hoban *et al.* 2017); it uses a similar user interface as
105 MINOTAUR. PHYLODYN provides functions to simulate phylogenies under a wide range of
coalescent models including heterogeneous sampling (Palacios *et al.* 2017). This package
also implements a variety of inference tools under Bayesian nonparametric coalecent.

108 Finally, Rousset *et al.* (2017) present the summary likelihood method of statistical
inference as an alternative to the approximate Bayesian computation (ABC) method,
which is also based on summary statistics when the full likelihood cannot be computed.
111 However, for the summary likelihood method of Rousset *et al.* the user does not need to
formulate priors on the distribution of the parameters. The package INFUSION provides a
generic implementation of this method, and Rousset *et al.* illustrate its use with a
114 coalescent model of population change.

Impacts of the hackathon and this special issue

The NESCent hackathon that gave rise to many of the tools and products reported in
117 this special issue was held with the kind of non-tangible objectives in nurturing the
community and fostering collaboration whose achievement will only truly manifest in the
long term. The collection of articles in this special issue will, hopefully, be a milestone

120 for the future progress of R in population genetics.

It is obviously too soon to pinpoint any such long-term successes. In addition to the tangible outcomes reported by Trainer *et al.* (2016) and the products described in this special issue, the hackathon nonetheless had a variety of impacts, some smaller and some larger. It helped create new collaborations on open source tools for population genetics, as evidenced by the author teams of the hackathon-related papers in this special issue; it introduced tools (such as the package HIERFSTAT) and their developers and users for the first time to collaborative code development and public version control on GitHub; and allowed participants to share knowledge and know-how. It also resulted in the revival of a previously existing but barely used community mailing list for population genetics in R (126 <https://stat.ethz.ch/mailman/listinfo/r-sig-genetics>). (129

It will remain to be seen whether these outcomes will have a lasting impact on population genetics as a field. However, the field will continue to face challenges – and opportunities – from the deluge of yet more massive amounts of genetic data, and these are likely to be addressed more effectively by a community well equipped to collaborate, whether across projects, institutions, or continents. (132 135

Open questions and future prospects

Population genetics is facing a number of challenges and opportunities from the next-generation sequencing revolution, both technical and scientific. On the technical side, how to efficiently deal with the massive amounts of data generated by next-generation sequencing will remain an ongoing problem. This includes the question of how data generated by different technologies are best analysed in combination. On the scientific side, the different drivers of genomic selection in different scenarios of population evolution have traditionally been investigated separately, using different genetic markers. Today's possibilities for genetic data collection offer the opportunity to assess how different portions of a genome are linked and evolve under different selective pressures in different environments. It will be particularly exciting to see how R and its packages will evolve to meet these challenges in the years to come. (138 141 144 147

Acknowledgements

We are grateful to the Editors of MER for accepting the idea of this special issue, the
150 authors of each article for their contribution, Armando Geraldes for his assistance, and
Shawn Narum and Karen Chambers for editorial support. The authors were participants
in the hackathon, and are indebted to NESCent (NSF #EF-0905606) for hosting and
153 supporting the event. This is publication ISEM 2016-236.

References

- Archer FI, Adams PE, Schneiders BB (2017) STRATAG: an R package for manipulating,
156 summarizing and analysing population genetic data. *Molecular Ecology Resources*,
THIS ISSUE, XXX–XXX.
- Danecek P, Auton A, Abecasis G, *et al.* (2011) The variant call format and VCFtools.
159 *Bioinformatics*, **27**, 2156–2158.
- Dray S, Dufour AB, Chessel D (2007) The ade4 package-II: two-table and *K*-table
methods. *R News*, **7**, 47–52.
- 162 Ellegren H (2014) Genome sequencing and population genomics in non-model organisms.
Trends in Ecology & Evolution, **29**, 51–63.
- Fisher RA (1930) *The genetical theory of natural selection*. Oxford University Press,
165 Oxford.
- Fordham DA, Brook BW, Moritz C, Nogués-Bravo D (2014) Better forecasts of range
dynamics using genetic data. *Trends in Ecology & Evolution*, **29**, 436–443.
- 168 Francis RM (2017) POPHELPER: an R package and web app to analyse and visualize
population structure. *Molecular Ecology Resources*, **THIS ISSUE**, XXX–XXX.
- Groen D, Calderhead B (2015) Science hackathons for developing interdisciplinary
171 research and collaborations. *eLife*, **4**, e09944.
- Hoban S, Archer F, DePrenger-Levin M, Liggins L, Parobek C, Strand A (2017)
SKELESIM: an extensible, general framework for population genetic simulation in R.
174 **THIS ISSUE**, XXX–XXX.
- Howison J, Herbsleb JD (2013) Incentives and integration in scientific software
production. *Proceedings of the 2013 Conference on Computer Supported Cooperative*

- 177 *Work*, pp. 459–470. ACM, New York.
- Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.
- 180 Jombart T (2008) *adeigenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.
- Jombart T, Archer F, Schliep K, *et al.* (2017) *apeX*: phylogenetics with multiple genes. *Molecular Ecology Resources*, **THIS ISSUE**, XXX–XXX.
- 183 Kamvar Z, López-Urbe MM, Coughlan S, Grünwald NJ, Lapp H, Manel S (2017) Developing educational resources for population genetics in R: an open and collaborative approach. *Molecular Ecology Resources*, **THIS ISSUE**, XXX–XXX.
- 186 Knaus BJ, Grünwald NJ (2017) VcfR: an R package to manipulate and visualize VCF format data. *Molecular Ecology Resources*, **THIS ISSUE**, XXX–XXX.
- 189 Lapp H, Bala S, Balhoff JP, *et al.* (2007) The 2006 NESCent phyloinformatics hackathon: a field report. *Evolutionary Bioinformatics Online*, **3**, 287–296.
- Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.
- 192 Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.
- 195 Luu K, Bazin E, Blum MGB (2017) *pcadapt*: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, **THIS ISSUE**, XXX–XXX.
- 198 Merilä J, Hendry AP (2014) Climate change, adaptation, and phenotypic plasticity: the problem and the evidence. *Ecological Applications*, **7**, 1–14.
- 201 Naisbitt J (1984) *Megatrends: ten new directions transforming our lives*. Warner Books, New York.
- Palacios JA, Karcher M, Minin VN, Lan S (2017) PHYLODYN: an R package for phylodynamic simulation and inference. **THIS ISSUE**, XXX–XXX.
- 204 Paradis E (2010) *pegas*: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, **26**, 419–420.
- 207 Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in

- R language. *Bioinformatics*, **20**, 289–290.
- Paradis E, Gosselin T, Goudet J, Jombart T, Schliep K (2017) Linking genomics and
210 population genetics with R. *Molecular Ecology Resources*, **THIS ISSUE**, XXX–XXX.
- Prins P, de Ligt J, Tarasov A, Jansen RC, Cuppen E, Bourne PE (2015) Toward
effective software solutions for big biology. *Nature Biotechnology*, **33**, 686–687.
- 213 Rousset F, Gouy A, Martinez-Almoyna C, Courtiol A (2017) The summary likelihood
method and its implementation in the INFUSION package. **THIS ISSUE**, XXX–XXX.
- Stanley RRE, Jeffery NW, Wringe BF, DiBacco C, Bradbury IR (2017) GENEPOEDIT: a
216 simple and flexible tool for manipulating multilocus molecular data in R. *Molecular
Ecology Resources*, **THIS ISSUE**, XXX–XXX.
- Trainer EH, Chaihirunkarn C, Kalyanasundaram A, Herbsleb JD (2014) Community
219 code engagements: summer of code & hackathons for community building in scientific
software. *Proceedings of the 18th International Conference on Supporting Group Work*,
GROUP '14, pp. 111–121. ACM, New York.
- 222 Trainer EH, Kalyanasundaram A, Chaihirunkarn C, Herbsleb JD (2016) How to
hackathon: socio-technical tradeoffs in brief, intensive collocation. *Proceedings of the
19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*,
225 CSCW '16, pp. 1118–1130. ACM, New York.
- Verity R, Collins C, Card DC, Schaal SM, Wang L, Lotterhos KE (2017) MINOTAUR: a
platform for the analysis and visualization of multivariate results from genome scans
228 with R Shiny. *Molecular Ecology Resources*, **THIS ISSUE**, XXX–XXX.
- Vitalis R, Gautier M, Klassmann A (2017) REHH 2.0: a reimplementation of the R
package REHH to detect positive selection from haplotype structure. **THIS ISSUE**,
231 XXX–XXX.
- Wringe BF, Stanley RRE, Jeffery NW, Anderson EC, Bradbury IR (2017)
parallelnewhybrid: an R package for the parallelization of hybrid detection using
234 NEWHYBRIDS. *Molecular Ecology Resources*, **THIS ISSUE**, XXX–XXX.

E.P. and T.G. co-edited this special issue. E.P., T.G., N.G., T.J., S.M., and H.L. conceived
and wrote the manuscript.
