Bioinformatics, 32(12), 2016, 1903–1904
doi: 10.1093/bioinformatics/btw098
Advance Access Publication Date: 18 February 2016
Applications Note

OXFORD

Genetics and population analysis

Coala: an R framework for coalescent simulation

Paul R. Staab and Dirk Metzler*

Department of Biology, Ludwig-Maximilians-Universität München, Planegg-Martinsried 82152, Germany

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on November 16, 2015; revised on February 6, 2016; accepted on February 14, 2016

Abstract

Summary: Simulation programs based on the coalescent efficiently generate genetic data according to a given model of evolution. We present *coala*, an R package for calling coalescent simulators with a unified syntax. It can execute simulations with several programs, calculate additional summary statistics and combine multiple simulations to create biologically more realistic data.
Availability and implementation: The package is publicly available on CRAN and on https://github.com/statgenlmu/coala under the conditions of the MIT license.
Contact: metzler@bio.lmu.de

1 Introduction

Coalescent simulators rapidly simulate the evolution of genetic sequences. They are routinely used in population genetics to approximate analytically intractable quantities, to verify theoretical predictions and to validate the performance of data analysis methods (Hoban *et al.*, 2012). The underlying idea to trace the ancestry of the sequences backwards in time was first implemented in the program *ms* (Hudson, 2002). Since then the size and complexity of data sets has increased, and many extensions and modifications of the original algorithm were proposed to cope with the new challenges. For example, the program *msHOT* (Hellenthal and Stephens, 2007) supports recombination hotspots, *MaCS* (Chen *et al.*, 2009), *fastsimcoal* (Excoffier and Foll, 2011) and *scrm* (Staab *et al.*, 2015) focus on the efficient simulation of long sequences, and *msms* (Ewing and Hermisson, 2010) and *cosi2* (Shlyakhter *et al.*, 2014) account for selection. The total number of published simulators is more extensive, and Peng *et al.* (2013) created an online database to help researchers select an appropriate program.

Currently, these simulators lack standardized in- and output formats and usually only support the calculation of a small number of summary statistics. To calculate additional statistics, researchers usually have to convert the simulated data into different formats. This process is error-prone and even small mistakes can lead to wrong conclusions or irreproducible results.

To simplify this process we present *coala*, an R package for calling coalescent simulators from within R. It provides a unified interface for specifying coalescent models in R (R Core Team, 2015), automatically selects a suitable simulator, transparently conducts the simulations, parses the results and calculates additional statistics from the output.

2 Implementation

The primary design goal of coala is extensibility. It uses a modular system based on R6 classes (https://cran.r-project.org/web/packages/R6) to integrate different simulators and summary statistics. Documentation for writing extensions is provided within the R package.

To set up a simulation study, the user creates a coala object called 'coalescent model' containing the demographic model and parameters to be used in the simulations as well as the specification which summary statistics shall be calculated from the simulated sequences. Parameters are variable values for quantities like times, rates or population sizes. Their values are either set when simulating the model later or are randomly sampled from a prior distribution. Optionally, parameters can be transformed using arbitrary R expressions including exponential and logarithmic functions. This design simplifies the execution of simulation-based analysis techniques such as ABC (Beaumont *et al.*, 2002) or Jaatha (Mathew *et al.*, 2013; Naduvilezhath *et al.*, 2011). An example ABC analysis using the *abc* package (Csilléry *et al.*, 2012) is included in the package's documentation.

Coala aims to work in a fully reproducible and transparent fashion. It displays the simulation commands that will be executed for a given model to allow the user to verify them. Additionally, it records all executed commands and ensures that all components respect R's

seeding system. An extensive suite of unit tests covers all aspects of the program.

Finally, the package is optimized for computational efficiency by using a buffering system combined with meta programming techniques for generating the simulation commands. The number of calls to the coalescent simulators is automatically minimized and computationally expensive parts are implemented in C++ (Stroustrup, 1995). Independent simulations can be distributed on multiple CPU cores.

3 Features

Although *coala* relies on existing programs for conducting the coalescent simulations, it can combine multiple programs to provide features not present in any of the applications. The current version of coala supports to call *ms* for standard neutral models, *msms* for models that contain selection and *scrm* for simulating long sequences. All three programs can be combined with *seq-gen* (Rambaut and Grassly, 1997) to simulate finite sites mutation models. Furthermore, *coala* automatically splits complex models into smaller parts that are supported by one of the coalescent simulators. It simulates the parts separately and combines the results. This can be used to simulate models that consist of genetic regions of different length or that have variable parameters between the regions. Finally, it can simulate unphased data by randomly mixing the phased haplotypes returned by the coalescent simulators and transcriptomic data by removing introns from the sequences.

Coala has support for calculating classical population genetic summary statistics like the nucleotide diversity π (Nei and Li, 1979), TajimaIN D (Tajima, 1989) and various kinds of frequency spectra. Additionally, the current version focuses on summary statistics that were developed for detecting selective sweeps. For instance, it utilizes the *rehh* package (Gautier and Vitalis, 2012) to calculate the EHH (Sabeti *et al.*, 2002), IES and iHS (Sabeti *et al.*, 2007) statistics and the program OmegaPlus (Alachiotis *et al.*, 2012) for the ω statistic (Kim and Nielsen, 2004). It can also return a basic SNP matrix, the ancestral trees in Newick format and the unmodified output of the simulators.

4 Conclusion

We have created an R package that simplifies the simulation of data according to evolutionary models. It allows researchers to conduct and process coalescent simulations in an easy, reliable and reproducible way. Different simulators can be called using a unified R syntax and a variety of summary statistics can be calculated from the results. We hope that coala will facilitate the use of more realistic models by allowing for variable mutation and recombination rates and finite sites mutation models. Although might not always be necessary to use sophisticated simulation models, the easy availability of these features facilitates testing whether they are important for a particular application.

Acknowledgements

We thank Ann Kathrin Huylmans for suggesting the name 'coala', and Soumya Ranganathan and the anonymous reviewers for helpful comments on the package's documentation.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft through the Research Unit FOR 1078 'Natural selection in structured populations' (DFG ME 3134/3-2) and the SPP Priority Programme 1590 'Probabilistic Structures in Evolution'.

Conflict of Interest: none declared.

References

Alachiotis,N. *et al.* (2012) OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*, **28**, 2274–2275.

Beaumont,M. *et al.* (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025

Chen,G.K. *et al.* (2009) Fast and Flexible Simulation of DNA Sequence Data. *Genome Res.*, **19**, 136–142.,

Csilléry,K. *et al.* (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.*, **3**, 475–479.

Ewing,G. and Hermisson,J. (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, **26**, 2064–2065.

Excoffier,L. and Foll,M. (2011) Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, **27**, 1332–1334.

Gautier,M. and Vitalis,R. (2012) rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, **28**, 1176–1177.

Hellenthal,G. and Stephens,M. (2007) msHOT: modifying Hudsonudsoncbi.nlm.nihto incorporate crossover and gene conversion hotspots. *Bioinformatics*, **23**, 520–521.

Hoban,S. *et al.* (2012) Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.*, **13**, 110–122.

Hudson,R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Kim,Y. and Nielsen,R. (2004) Linkage disequilibrium as a signature of selective sweeps. *Genetics*, **167**, 1513–1524.

Mathew,L.A. *et al.* (2013) Why to account for finite sites in population genetic studies and how to do this with Jaatha 2.0. *Ecol. Evol.*,

Naduvilezhath,L. *et al.* (2011) Jaatha: a fast composite likelihood approach to estimate demographic parameters. *Mol. Ecol.*, **20**, 2709–2723.

Nei,M. and Li,W.H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci U S A*, **76**, 5269–5273.

Peng. *et al.* (2013) Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators. *Bioinformatics*, **29**, 1101–1102.

Rambaut,A. and Grassly,N.C. (1997) Seq-Gen: An Application for the Monte Carlo Simulation of DNA Sequence Evolution Along Phylogenetic Trees. *Comput. Appl. Biosci*, **13**, 235–238.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sabeti,P.C. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Sabeti,P.C. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Shlyakhter,I. *et al.* (2014) Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, **30**, 3427–3429.

Staab,P.R. *et al.* (2015) scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, **31**, 1680–1682.

Stroustrup,B. (1995). *The C++ Programming Language*, 3rd edn. Pearson Education India.

Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.