

MOLECULAR ECOLOGY RESOURCES

SkeleSim: an extensible, general framework for population genetic simulation in R

Journal:	<i>Molecular Ecology Resources</i>
Manuscript ID	Draft
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Parobek, Christian; University of North Carolina at Chapel Hill, Curriculum in Genetics and Molecular Biology Archer, Frederick; Southwest Fisheries Science Center, DePrenger-Levin, Michelle; Denver Botanic Gardens Hoban, Sean; Morton Arboretum Liggins, Libby; The University of Queensland, School of Biological Sciences Strand, Allan; College of Charleston, Biology; College of Charleston, Grice Marine Laboratory
Keywords:	Conservation Genetics, Contemporary Evolution, Ecological Genetics, Landscape Genetics, Molecular Evolution, Wildlife Management
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
skeleSim Flow.svg	



Which scenario

2

Add a new scenario

Number of Populations

2

Migration Model

user

Migration rate multiplier (no effect for model 'user')

1

Type of locus

microsatellite

Number of loci

20

Population characteristics

Among population migration

Locus characteristics

Columns represent *from* and rows represent *to*. Migration models determine the structure of the matrix. The migration multiplier changes the elements in the matrix by this factor (direct multiplication). To enter arbitrary elements in the matrix choose the "user" migration model. Spatial arrangements (twoD, twoDwDiagonal, and distance) require that the populations be arranged in rows and columns. The product of the rows and columns must equal the number of populations. In addition, twoD and twoDwDiagonal have to have both rows and columns > 1.

Migration

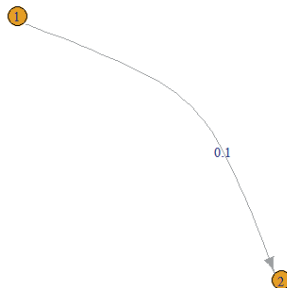
matrix number

0

1 migration matrix defined currently
(indexed from '0')

Migration Matrix

0	0
0.1	0



skeleSim

Actions

Help Choosing Simulator

General Conf

Scenario Conf

Rmetasim Params

Results

Current ssClass

Type of plots

- ☐ Summarize Scenario
- ☒ Compare Scenarios

Global statistics

Chi2 Chi2.pval F'st F'st.pval Fis Fis.pval Fst
Fst.pval G'st G'st.pval G'st G'st.pval Gst
Gst.pval

Locus-level statistics

allelic.richness exptd.heterozygosity Fis Fit Fst
hwe.p mRatio num.alleles num.priv.allele
obsvd.heterozygosity prop.genotyped
prop.unique.alleles theta


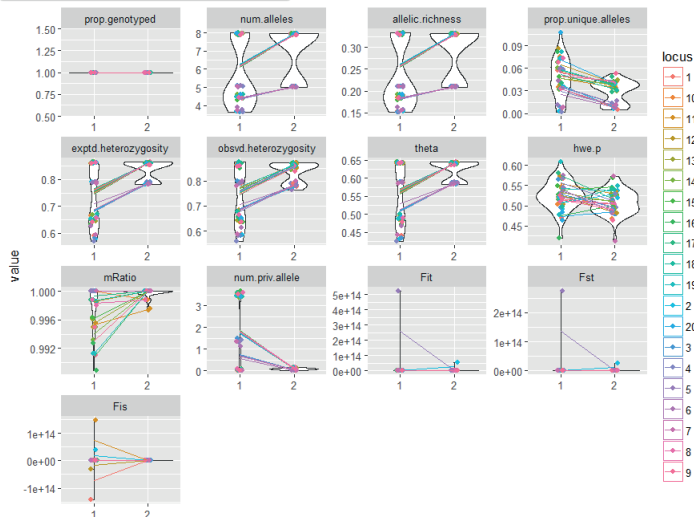
Pairwise statistics

Chi2 Chi2.pval chord.dist F'st F'st.pval Fis
Fis.pval Fst Fst.pval G'st G'st.pval G'st
G'st.pval Gst Gst.pval shared.alleles

Global Statistics

Locus Statistics

Pairwise Statistics

 Download .csv of locus analyses


Archer FI, Adams PE, Schneiders B (201X) strataG: An R package for manipulating, summarizing, and visualizing phylogenetic data. *Molecular Ecology Resources* 19: 1–12.

Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. Models and estimation procedures. *Annual Review of Ecology and Systematics* 8: 285–310.

Garza JC, Williamson EG (2001) Detection of reduction in population size using data from microsatellites. *Molecular Ecology* 10: 85–95.

Fu Y-X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Molecular Biology and Evolution* 14: 416–426.

Goudet J (2005) hierfstat, a package for R to compute and test hierarchical F -statistics. *Molecular Ecology Resources* 5: 149–150.

Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, Oxford. (doi:10.1093/oxfordjournals.mol.a026316)

Paradis E (2010) pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26: 1369–1370.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 137: 1315–1332.

Takezaki N, Nei M (1996) Genetic distances and reconstruction of Phylogenetic trees from microsatellite data. *Molecular Biology and Evolution* 13: 125–132.

Weir BS, Cockerham CC (1984) Estimating F -statistics for the analysis of population structure. *Evolution* 38: 136–140.

Weir BS, Hill WG (2002) Estimating F -statistics. *Annual Review of Genetics* 36: 721–750.

For Review Only

and analyzing population genetic data. *Molecular Ecology Resources*
American Journal of Human Genetics 19 (3 Pt 1): 233.
Nullite loci. *Molecular Ecology* 10: 305-318.
Background selection. *Genetics* 147:915-925.
Molecular Ecology Notes 5: 184-186.
: pp. 256, eqn 12.67)
Bioinformatics 26: 419-420.
Genetics 123:585-595.
Nullite DNA. *Genetics* 144:389-399.
Genetics 38: 1358-1370.

For Review Only

skeleSim: an extensible, general framework for population genetic simulation in R

Christian M. Parobek¹, Frederick I. Archer², Michelle E. DePrenger-Levin³, Sean M. Hoban⁴, Libby Liggins⁵, and Allan E. Strand⁶

¹Curriculum in Genetics and Molecular Biology, University of North Carolina, Chapel Hill, USA.

²Southwest Fisheries Science Center, 8901 La Jolla Shores Drive, La Jolla, CA 92037

³Denver Botanic Gardens, 909 York Street, Denver, CO 80206

⁴Morton Arboretum, 4100 Illinois Route 53, Lisle, IL 60532

⁵Institute of Natural and Mathematical Sciences, Massey University, Auckland 0745, New Zealand

⁶College of Charleston, 66 George Street, Charleston, SC 29424 USA

Running title: skeleSim: population genetic simulation in R

Keywords

population genetics, simulations, the coalescent, forward-time, conservation genetics, open-source, null model, power-analysis

ABSTRACT (<250 words)

Simulations are a key tool in molecular ecology for inference and forecasting, as well as for evaluating new methods. Due to growing computational power and a diversity of software with different capabilities, simulations are becoming increasingly powerful and useful. However, the widespread use of simulations by most geneticists and ecologists is hindered by difficulties in understanding these software's complex capabilities, composing code and input files, a daunting bioinformatics barrier, and a steep conceptual learning curve. skeleSim (an R package) guides users in choosing appropriate simulations, setting parameters, calculating summary genetic statistics, and organizing data output, in a reproducible pipeline within the R environment. skeleSim is designed to be an extensible framework that can 'wrap' around any simulation software (inside or outside the R environment) and be extended to calculate and graph any summary genetic statistics. Currently, skeleSim implements coalescent and forward-time models available in the fastsimcoal2 and rmetasim simulation engines to produce null distributions for multiple population genetic statistics and marker types, under a variety of demographic conditions. skeleSim is intended to make simulations easier while still allowing full model complexity to ensure that simulations play a fundamental part of molecular ecology investigations. skeleSim can also serve as a teaching tool: demonstrating the outcomes of stochastic population-genetic processes; general concepts of simulations; and providing an introduction to the R environment with a user-friendly graphical user interface (using 'shiny').

INTRODUCTION

Simulations of genetic and environmental processes have diverse uses in ecology and evolutionary biology research (Hoban 2014), as well as applications in agriculture and aquaculture, public health and conservation (e.g. fishery forecasts, pathogen evolution, and extinction risk). In the past decade, simulations have been increasingly popular for inferring the historical processes that resulted in current patterns in molecular data (Marino *et al.* 2013; Jombart *et al.* 2014), predicting the molecular genetic outcomes of complex future processes (Hedrick 1995; Bruford *et al.* 2010), and evaluating methods (Girod *et al.* 2011; Hoban *et al.* 2013a) and sampling strategies (Oyler-McCance *et al.* 2013; Lotterhos & Whitlock 2015). To generalize, simulations are used to create many *in silico* genetic datasets of individuals and populations which could have been produced under a given model of a real system. Someone using simulations will often wish to model a range of scenarios, such as the different degrees of hybridization, or different population or species' divergence times. Summaries of datasets generated under these scenarios can then be compared to quantitatively establish which model is most consistent with real data, to generate hypotheses or predictions, to explore model sensitivity to particular parameters (e.g. population sizes), or to decide on an appropriate sampling strategy, study methods, or management approach.

Dozens of software packages that implement simulations of demography, ecology, genetics, spatial processes, behavior, adaptation, interspecies interactions, and more now exist (Hoban *et al.* 2011; Peng *et al.* 2013), allowing scientists to use available packages rather than code their own simulation software from scratch. Existing simulation software varies in complexity (Carvajal-Rodriguez & Antonio 2008; Hoban *et al.* 2011), providing a wide range of options that can make many simulators highly useful and flexible. Flexibility, however, often comes at a cost;

many simulators require substantial investment in learning complex user interfaces and commands, the preparation of custom code and input files, and in-depth immersion and experimentation to explore suitable model space. Also required for the use of any simulator are relatively strong bioinformatics skills to prepare a series of simulation scenarios, produce many genetic datasets, analyze the data for various genetic summary statistics, and organize this output. Despite an increasing number of tutorials, books, workshops, and articles aimed at bioinformatic training for biologists (Haddock & Dunn 2011; Münkemüller *et al.* 2012), acquiring the knowledge and skills necessary to use simulation software continues to be a barrier for many potential users.

A user-friendly interface and analysis pipeline that guides a user through the steps of setting up a model, choosing analyses, running a simulation, and visualizing results would help circumvent obstacles that prevent population geneticists from using simulations in their research. Although several tools have made progress towards this goal, each has limitations. For instance: *MODELER4SIMCOAL2* (Antao *et al.* 2007) provides a graphical interface to help write simulation input files for *simcoal*, including complex demographies; and *PopPlanner* (Ewing & Hermisson 2010) is a graphical tool which can be used to construct *ms* and *msms* command lines that model various scenarios. The downside of these softwares, however, is that they are specific to only one simulator, and they only construct the simulations and do not organize or analyze the datafiles. Other examples of user-friendly simulation softwares include: *SPOTG* (Hoban *et al.* 2013b) and *PowSim* (Ryman *et al.* 2006) which are graphical and command-line interfaces that perform simulations and calculate statistical power of different sampling strategies; and *onesamp* (Tallmon *et al.* 2008) that uses coalescent simulations to infer effective population size (N_e). These softwares help users analyze datafiles, however each is

designed for a specific use of simulations, and therefore restrict users to certain scenarios and summary statistics. In contrast, *coala* (Staab & Metzler 2016) is an R package that can wrap several coalescent simulators, standardising the input and output files across softwares, and offers the calculation of summary statistics. Such a software-model that enables users to apply learned skills across different softwares is very desirable. In addition, such a software would ideally be built in an extensible, flexible way to enable the use of any simulator (both coalescent and forward-time), for many applications of simulations, and for a variety of current and potentially future genetic analyses.

We introduce *skeleSim*, a new R package that will help molecular ecologists create and use simulations for a wide variety of purposes. We have aimed for maximum flexibility, power, guidance, and user-friendliness. We envision that this software will be used in teaching, research, and applied-science situations, such as biodiversity conservation and management. We implement our software in R because it is freely available, works on all operating systems, has powerful statistical and graphing functions, and is open-source, allowing users to access, modify, and extend code and package capability. R is commonly used in ecology, evolution, and epidemiology across all stages of career development because it offers a supportive learning environment (with provisioning of vignettes and tutorials) and an active and responsive community. These features of R, and a recently developed user-friendly graphical interface, called ‘shiny’ (Chang *et al.* 2015), will enable molecular ecologists to make use of, and learn with, *skeleSim* regardless of their coding ability.

skeleSim is essentially a ‘control panel’ or ‘wrapper’ to enable the use of existing simulation software, without need for the user to create input files or write an informatics pipeline. The

process is summarized as follows. A user enters parameters such as population sizes and migration rates, and makes choices regarding demographic events. `skeleSim` will take these choices and create input files or input code, and then simulate the various scenarios requested by the user. Next, `skeleSim` will calculate a suite of summary genetic statistics and graphically display these according to user-specified choices. The current version encompasses most widely used statistics in population genetics, with the capacity to add many more including user-defined statistics. `skeleSim` will also organize the results into a convenient list object in R, so that statistics, replicates, and scenarios can be easily accessed, subsetted, analyzed, and summarized.

`skeleSim` has two desirable features that do not exist in current simulation softwares. First, `skeleSim` implements a series of automatic validation steps at multiple stages to ensure smooth operation of the simulation engine once the user is ready to run replicates (**Figure 1**). Classic problems that users of simulation software experience include: small formatting or text errors, incorrect file paths, missing files, and parameters that are incompatible or prevent convergence, all of which can cause the software to crash. These issues take time to identify and resolve. Few simulation software have such extensive internal controls to ensure the user has entered valid parameters to create feasible and realistic models with the possibility of coalescence within basic time and memory limits. Second, we have created an R S4 class for each simulation that will store data, results, and the parameters of all scenarios; facilitating complete documentation of all simulations run. Thus the parameters are permanently attached to the data itself without the need for outside documentation, which is more easily lost or corrupted. The parameters, the data, and the software for running new simulations are thus easily shareable. This feature also enables easy ‘tweaks’ to scenarios that have been run, such

as to explore a new parameter value while keeping all other aspects of the simulation unchanged.

skeleSim was first envisioned at the R PopGen Hackathon, sponsored by the National Evolutionary Synthesis Center (NESCent), Durham, North Carolina, USA (as will likely be described in the introduction to the special issue). The need for an accessible framework for using simulations was identified by a group of over 20 population geneticists. This team helped scope and design the main features and uses of skeleSim.

MATERIALS AND METHODS

Installation:

skeleSim is organized as a standard R package such that normal installation of the package will also install most packages on which it depends, such as strataG, apex, adegenet, pegas, and rmetasim. fastsimcoal2 is a separate executable which must be manually installed and placed in the appropriate directory. Links and instructions are provided in the 'Installing' vignette in skeleSim.

Supported engines:

skeleSim can interface with simulation engines written either as native R packages (eg. rmetasim), or pre-compiled system executables (eg. fastsimcoal2), making it flexible and extensible. Currently, we have built skeleSim to support a forward-time simulator (metasim, (Strand 2002)) and a coalescent simulator (fastsimcoal2, (Excoffier & Foll 2011)). Metasim and fastsimcoal2 are two of the most powerful and widely used simulation software available, and enable complex simulations (**Table 1**). Both allow simulation of a

multiple-population system of arbitrary population sizes and migration rates, through potentially long periods of time (tens to thousands of generations). Also, both softwares allow multiple types of genetic markers for which the user specifies mutation rates (e.g. sequence data, microsatellite, and SNP data). *Metasim* is especially advantageous because it allows for simulating complex life history and demography, such as multiple life history stages and stage / age specific mortality and reproduction rates. *Fastsimcoal2* is well suited for complex series of events in history such as bottlenecks, divergence events, and admixture among populations, and allows realistic genomic features such as linkage among markers, long sequences, and recombination rates. Though both softwares have detailed user guides to help the user navigate their high flexibility and detailed code, there is still a steep learning curve for these software, making them especially suitable for *skeleSim*. *skeleSim* helps users to access the functionality of *metasim* and *fastsimcoal2* with little prior knowledge of the features and parameter options they provide. Furthermore, the *skeleSim* wrapper functions allow more advanced users to construct, run, and analyze complex scenarios with a reproducible pipeline. Thus, *skeleSim* caters to the wide range of skill levels found within the molecular ecology community, and facilitates and accommodates skill development in the course of a molecular ecologist's career.

Choosing between coalescent and forward-time simulation:

The user can either directly select a forward-time or coalescent simulation or allow the software to provide guidance on simulator selection. Generally, simulators are classified into one of two categories: coalescent (or backward) and forward-time (Hoban *et al.* 2011). Most forward-time simulations are individual-based, while coalescent simulators follow genetic lineages backwards in time. As a broad rule, coalescent simulators are suited for organisms with simple life histories,

but complex demographic histories (e.g. multiple population divergences, complex changes in population size) and large populations (large meaning effective size of tens of thousands). Forward-time simulations are best suited for organisms in which life history is important (e.g. variance in reproduction among individuals, age structure, age-based migration) and large population sizes and long-time scales are not needed. An additional difference is computational speed: coalescent simulators can produce replicate simulations much faster than forward-time simulators and are typically preferred when complex demography is not required in a simulation. Note that there are also hybrid simulators such as the recent MetaPopGen (Andrello & Manel 2015), which are forward-time but follows genotype frequencies rather than individuals. The distinctions among simulation categories are discussed in detail elsewhere (Carvajal-Rodriguez & Antonio 2008; Liu *et al.* 2008; Hoban *et al.* 2011).

Interface:

`skeleSim` is designed primarily to be used in the shiny graphical user interface. Guidance on installing `skeleSim` and calling `skeleSimGUI()` is provided in the `skeleSim` vignette 'Installing'. The vignette 'Running simulations' provides an overview of the steps and describes the processes, labelling, and construction of files that happens 'behind the scenes' of the interface including some basic troubleshooting. The interface itself is constructed to be sequential tabs, each with their own description that guides a user through necessary steps. First, the user may choose whether they wish to run `fastsimcoal2` or `rmetasim`, or the user may receive guidance on this choice through a series of questions ('Help Choosing Simulator' tab, e.g. does the simulation require an organism with complex life history, what are the computational and time limits of the user's system). Next, the user defines: general parameters in the 'General Conf' tab, including a title and selecting types of summary genetic statistics;

scenarios in the ‘Scenario Conf’ tab, including the number of study populations and the type of locus; and simulator-specific parameters (e.g. either ‘Rmetasim Params’ tab or ‘FastSimCoal params’ tab currently). In each case, the `skeLeSim` interface presents labelled text boxes and drop-down menus of the required and optional parameters (with some further explanation). As it is common practice to change one parameter per scenario for comparisons, a user can define additional scenarios by simply modifying the first scenario. A given study may examine two or up to dozens of scenarios, depending on its complexity and purposes (see (Hoban *et al.* 2011) for more guidance on designing scenarios). The parameters of each scenario are saved and the user is able to run the simulation. The second to last tab of the `skeLeSim` interface is ‘Results’. In this tab, users can upload their simulation output results and quickly visualize the summary genetic statistics for each scenario within their simulation and compare results among scenarios. The last tab of the interface is ‘Current ssClass’. This tab allows the user to visualise how the `skeLeSim` S4 class object in R is altered by options and operations executed within the interface, helping to familiarize the naive R user with coding conventions.

Architecture:

All parameters of each scenario and results (including full output of all replicates, and all summary genetic statistics) are contained in a single S4 class object. Users parameterize the object using the `shiny` web browser interface, or for more experienced users, directly via R code. This object can be saved at any time and re-loaded in the `shiny` interface, or in any R environment (eg. to run later on a different personal computer or a server).

All primary functions in `skeLeSim` receive this S4 class object as their single argument, thus providing the function with all information about the simulation. Functions can also add

information to this object in predefined slots and return modified versions of the object to the workspace. In this manner, the course of the simulations, from parameter specification to simulation output and analysis, is fully captured. This ensures that the results of analyses will be permanently linked to the parameters and models used to produce the data, a relatively novel feature in simulation software.

Summary statistics:

We have included numerous summary genetic statistics within `skeleSim` to describe simulation outputs. To help guide the selection of appropriate summary statistics to consider, users are presented with suites of summary statistics nested under categories that align with hypotheses relating to: alpha diversity or population-specific measures ('Locus Statistics', e.g. number of alleles, m -ratio), beta diversity or population-pairwise measures ('Pairwise Statistics', e.g. F_{ST} , nucleotide divergence), and global measures ('Global Statistics', e.g. global F_{ST}).

Analysis options can be customised by advanced users, by nesting further summary statistic options under the existing categories, or creating a new category. Routines for calculating population genetic statistics are sourced from existing R packages including `strataG` (Archer *et al.* Submitted to Molecular Ecology, 2016) and `adegenet` (Jombart 2008) that offer interoperability and complementary analysis options for population geneticists. A full list of summary genetic statistics available in the current version of `skeleSim` are described in

Appendix 1.

Forecasting:

The 'Forecasting' vignette demonstrates how simulations may be used to forecast possible outcomes of rare-species management. Conservation managers are often faced with decisions

about corridors or translocations to link two populations. A user may want to implement a simulation in which two populations of different sizes are: disconnected (Scenario 1); or connected via gene flow (Scenario 2). Using the `skeleSim` interface, the two scenarios are constructed by the user to differ only by an asymmetric migration rate (0.10) from the larger population to the smaller population in Scenario 2, to mimic translocation by conservation managers (**Figure 2**). These scenario parameters can be saved and the simulations for both can then be executed simultaneously.

In this example, where one population is quite small and the other large, the user may be interested in whether this level of gene flow (Scenario 2) sufficiently maintains genetic diversity and counteracts drift in the small population, and whether the small population's unique diversity is swamped by gene flow from the larger population. The results tab of the `skeleSim` interface allows the user to immediately compare results among scenarios. In this example, the user will quickly see that by comparing the 'Locus statistics' for Scenario 1 and 2, that the smaller population has a greater number of alleles (`num.alleles`), higher allelic richness (`allelic.richness`), and observed heterozygosity (`obsvd.heterozygosity`) in Scenario 2, where there is migration from the larger population (**Figure 3**; see **Appendix 1** for further explanation of analyses). However, by observing the 'Global statistics' the users will also see that, intuitively, the global population structure (e.g. F_{ST}) is reduced in Scenario 2, as a consequence of both the proportion of unique alleles (`prop.unique.alleles` in 'Locus Statistics') in the smaller population being reduced by gene flow, and the number of private alleles (`num.priv.alleles` in 'Locus Statistics') found in Population 1 also being decreased in Scenario 2. Further examples are provided as vignettes.

DISCUSSION

skeLeSim occupies a unique niche that has long been neglected in simulation software, which is a user-friendly, streamlined interface for the entire process of *using* the simulations, from setting up the models, documenting pipelines, to obtaining organized results. It is often difficult for a user to know which parameters are required, to determine whether their code or input files will run successfully, and to interpret error messages. While some simulation software incorporate built-in analyses (e.g. *metasim*, *cdpop*), and some have visually accessible interfaces (e.g. *Pedagog*), few software have both, and none have been designed and structured in order to facilitate the entire process of using simulations (perhaps with the exception of some Approximate Bayesian Computation packages). Many simulation software are stand-alone programs, which adds an additional step of organizing and importing data to other statistical software, such as R, where one can use statistical and graphical approaches to draw inference from the simulations.

skeLeSim will lower the initial time and knowledge needed to start doing simulations. By helping to bring genetic simulators to a wider audience, we ultimately hope simulation tools will be more widely used in ecology and evolutionary biology. Simulations complement and strengthen empirical investigations at multiple stages, from planning a study to interpreting results to applying models and data in a predictive fashion for forecasting (Hoban 2014). Their use will enable greater power and rigor of studies in molecular ecology (Epperson *et al.* 2010; Balkenhol & Landguth 2011; Andrew *et al.* 2013). We also see skeLeSim as a platform that will make simulations more usable to groups such as conservation practitioners, scientists in public health and agriculture, and educators.

`skeleSim` is designed to be extended to new situations and simulators. Outside developers can add wrappers for new simulators following the examples set by the wrappers for `fastsimcoal` and `rmetasim`. That being said, `skeleSim` will be most easily extensible to other simulation software that have similar models and parameters to those that we have implemented, such as `kernelpop`, `cdpop`, and `nemo`, and coalescent software like `ms`. Additional analyses are also able to be implemented as updates to `skeleSim` by external developers. For example, currently linkage disequilibrium and estimates of effective population size are not implemented as these analyses are computationally intensive and not as suitable for large-scale simulations. Nonetheless, these may be important outcomes in some studies. The analytical result slot of the `skeleSim` S4 class object is an R list that is currently structured to hold summary statistics for each population, pairs of populations, and globally (i.e. all populations). New statistics that fit one of those categories can be easily added to the current analytical functions or a new category can be created to store summaries that do not fit these categories, such as linkage disequilibrium.

One aspect not implemented in `skeleSim` is natural selection. We chose not to do so at this time for several reasons. Implementing natural selection would greatly increase the potential range of options for the user. A common use of simulators is to generate 'null' distributions of statistics - the statistics that we expect to occur as a product of the neutral processes of mutation, drift and migration. These null distributions have immense utility for inference of demographics and also potentially for identifying 'outlier' loci that do not fall within such distributions. Nevertheless, future work on `skeleSim` may implement selection.

skeleSim is a dynamic package that we hope will grow in capability based on feedback from users and additions from other developers. We have concentrated initial development of the package on helping users to produce null distributions of statistics under scenarios based on varying the most common demographic parameters. In order to rapidly develop this proof-of-concept package, we have not incorporated some useful features, such as introducing genotyping or sequencing error into the simulated data, analyzing empirical user-contributed data alongside simulated data, or allowing a user to specify starting conditions for the simulations, such as the distribution of genotype frequencies. Finally, in addition to generating null distributions, we envision expanding skeleSim to include modules to examine the statistical power of various tests as well as to conduct performance testing of analytical methods.

Recently developed tools in molecular ecology are user friendly, which is important due to the many and complex tools that allow precise and sound conclusions. For example, (Gruber & Adamack 2015) recently released LandGenReport, a comprehensive R package that for the first time implements the multiple steps of landscape genetic analysis (see (Segelbacher *et al.* 2010) or (Manel *et al.* 2003) for an overview of landscape genetics) in one framework (Manel *et al.* 2003; Segelbacher *et al.* 2010; Gruber & Adamack 2015). Similarly, Frichot and Francois recently created LEA to assist scientists in performing gene-environment association studies with genome scans (*sensu* (Eckert *et al.* 2010; Coop *et al.* 2010)), to identify loci that could underlie local adaptation (Frichot & François 2015). As the field of molecular ecology expands, skeleSim fulfills a need of making software tools and analyses extremely accessible to a wide audience. Users are encouraged to fork the skeleSim code from GitHub and suggest or contribute to updates, new analyses, and new simulators for skeleSim.

Acknowledgements

The resource reported in this paper originated at the Population Genetics in R Hackathon, which was held in March 2015 at the National Evolutionary Synthesis Center (NESCent) in Durham, NC, with the goal of addressing interoperability, scalability, and workflow challenges for the population genetics package ecosystem in R. The authors were participants in the hackathon, and are indebted to the event organizers (T. Jombart, S. Manel, E. Paradis, and H. Lapp), other participants, and NESCent (NSF #EF-0905606) for hosting and supporting the event. Ongoing development of this resource was supported by the National Institute of Mathematical and Biological Synthesis (NIMBioS) through a funded short-term visit. CMP was supported by funding from the NIH: T32GM007092 and F30AI109979. LL was supported by an Allan Wilson Centre for Molecular Ecology and Evolution Postdoctoral Fellowship and a Rutherford Foundation New Zealand Postdoctoral Fellowship.

Data Accessibility

- Source code and current development version available from GitHub (github.com/christianparobek/skeleSim)
- Vignettes ship with the package, and additionally can be accessed at (github.com/christianparobek/skeleSim/vignettes)
- Stable version of `skeleSim` and vignettes will be available from CRAN

Author Contributions

- All authors contributed to the idea conception, coding, testing, and manuscript writing.

Table 1: A comparison of the functionality of the simulation software accessed by skeleSim, SimCoal and Metasim.

Aspect	Metasim	Simcoal
Algorithm focused on:	Individuals	Lineages
Population	Population and individual based, migration controlled by migration matrix, migration can be different for each sex or stage	Population and sample based, migration controlled by migration matrix
Lifecycle / Stages	User defined number of stages, each with user defined survival rates	Single stage, Wright Fisher (all individuals live for one time step)
Population growth rate	Arises through reproduction matrices, up to a hard carrying capacity at which point individuals are removed at random	User defines exponential growth rate (positive or negative)
Mating	Random mating within population; proportion selfing?	Random mating within population
Migration	Movement of either male gametes or offspring	Movement of proportion of the population adults
“Events” allowed	Change in migration rates, demographic matrices; harder to code, and not currently implemented in Skelesim	Population fission/ fusion, change in population sizes & migration rates; very easy to change
Mutation rate	Sequence-wide mutation rate, stepwise mutation model for microsatellites	Substitution rate for sequences, stepwise mutation model for microsatellites
Recombination among loci	No	Yes
Natural selection	None	None
Other features of interest	Tracks previous generation pedigree (parent/ offspring relations) and population of origin for individuals	Can define linkage between markers and thus construct chromosomal segments

Figure 1: An outline of the steps required to undertake a population-genetic simulation study, and the sources of information and support available to the user when using *skeleSim* versus the standalone *metasim* or *fastsimcoal2* software. Using the R platform, *skeleSim* provides centralized support for the user throughout the decision-making and technical steps involved in a population-genetic simulation study. Furthermore, *skeleSim* enables the validity of a population-genetic scenario to be checked early (indicated by * and retrograde arrows), avoiding considerable time investment in unnecessary data formatting, analysis and visualization.

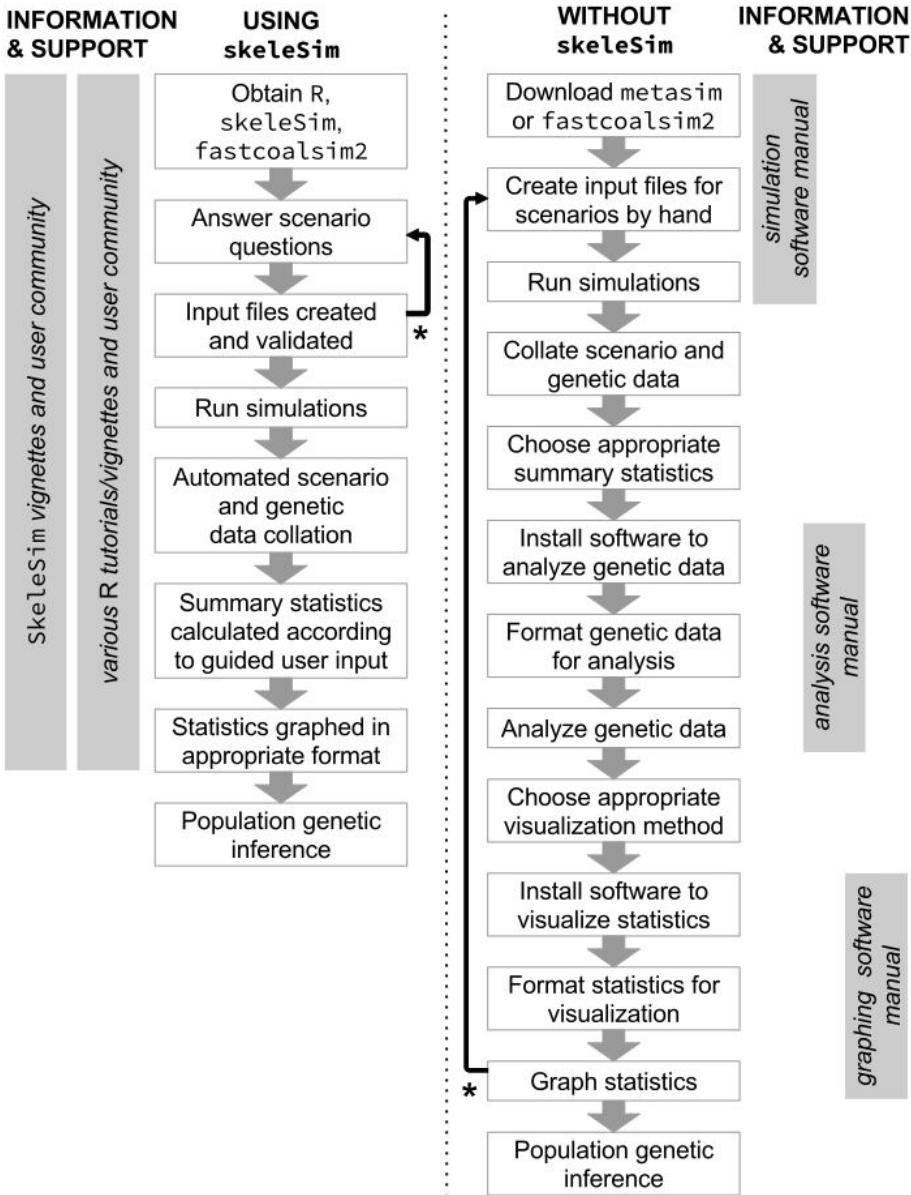


Figure 2: The 'Scenario Conf' tab of the skeleSim user interface. In this tab the simulation scenarios are defined by the user. Migration rate and directionality are defined by the user in the matrix, and a matching population graph is automatically populated in the interface. This population graph corresponds to Scenario 2 of the 'Forecasting' example (see main text and skeleSim vignettes).

skeleSim
Actions
Help Choosing Simulator
General Conf
Scenario Conf
Rmetasim Params
Results
Current ssClass

Which scenario
2
Add a new scenario

Number of Populations
2

Migration Model
user

Migration rate multiplier (no effect for model 'user')
1

Type of locus
microsatellite

Number of loci
20

Population characteristics
Among population migration
Locus characteristics

Columns represent *from* and rows represent *to*. Migration models determine the structure of the matrix. The migration multiplier changes the elements in the matrix by this factor (direct multiplication). To enter arbitrary elements in the matrix choose the "user" migration model. Spatial arrangements (twoD, twoDwDiagonal, and distance) require that the populations be arranged in rows and columns. The product of the rows and columns must equal the number of populations. In addition, twoD and twoDwDiagonal have to have both rows and columns > 1.

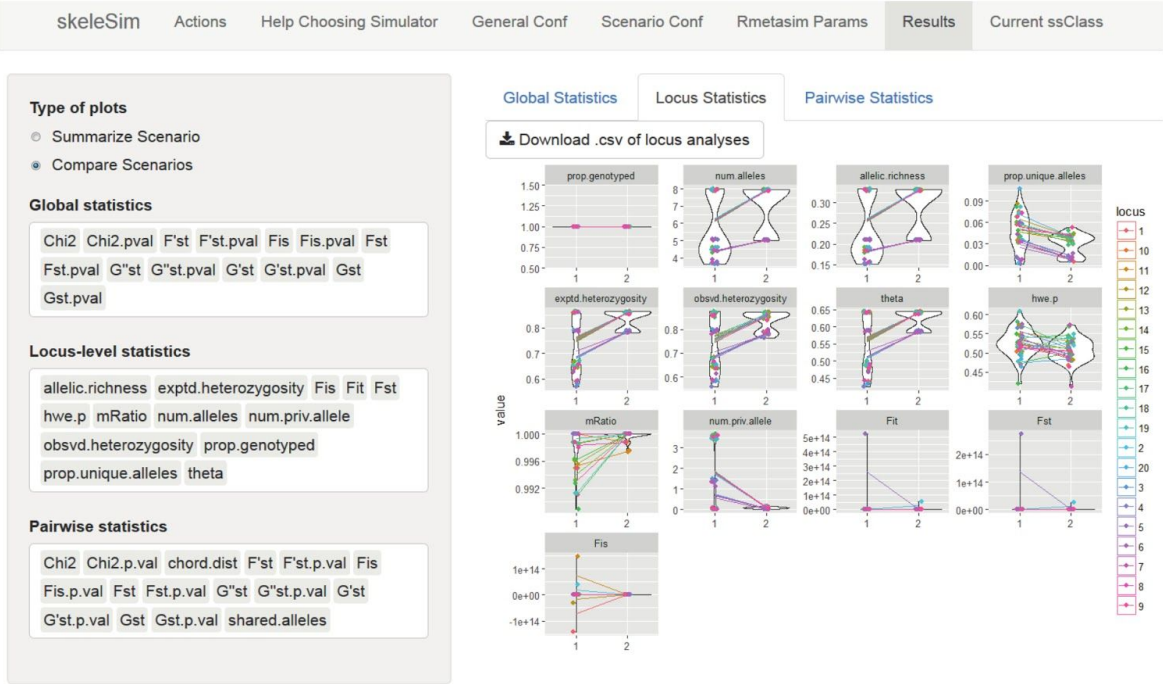
Migration matrix number
0

1 migration matrix defined currently (indexed from '0')

Migration Matrix

0	0
0.1	0

Figure 3: The ‘Results’ tab of the skeleSim user interface. Following the completion of simulations, users can upload and visualize the results for the genetic summary statistics they elected. In this ‘forecasting’ example (see main text and skeleSim vignettes), the results for the ‘Locus-level statistics’ from having no migration among populations (Scenario 1) and migration from the larger population to the smaller population (Scenario 2) can be observed. The rapid visualization of simulation results enabled by skeleSim facilitates prompt decision-making for conservation, and any subsequent scenario modifications if necessary.



Appendix 1: Overview of the genetic summary statistics available in the current version of skeleSim. Analyses are nested within the categories 'Locus', 'Global', and 'Pairwise'. skeleSim accesses functions within existing R packages.

REFERENCES

- Andrello M, Manel S (2015) MetaPopGen: an r package to simulate population genetics in large size metapopulations. *Molecular ecology resources*, **15**, 1153–1162.
- Andrew RL, Bernatchez L, Bonin A *et al.* (2013) A road map for molecular ecology. *Molecular ecology*, **22**, 2605–2626.
- Antao T, Beja-Pereira A, Luikart G (2007) MODELER4SIMCOAL2: a user-friendly, extensible modeler of demography and linked loci for coalescent simulations. *Bioinformatics*, **23**, 1848–1850.
- Archer FI, Adams PE, Schneiders B (Submitted to Molecular Ecology, 2016) strataG: An R package for manipulating, summarizing, and analyzing population genetic data.
- Balkenhol N, Landguth EL (2011) Simulation modelling in landscape genetics: on the need to go further. *Molecular ecology*, **20**, 667–670.
- Bruford MW, Ancrenaz M, Chikhi L *et al.* (2010) Projecting genetic diversity and population viability for the fragmented orang-utan population in the Kinabatangan floodplain, Sabah, Malaysia. *Endangered species research*, **12**, 249–261.
- Carvajal-Rodriguez A, Antonio C-R (2008) Simulation of Genomes: A Review. *Current genomics*, **9**, 155–159.
- Chang W, Cheng J, Allaire J, Xie Y, McPherson J (2015) *Shiny: web application framework for R*.
- Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.
- Eckert AJ, Liechty JD, Tarse BR, Pande B, Neale DB (2010) DnaSAM: Software to perform neutrality testing for large datasets with complex null models. *Molecular ecology resources*, **10**, 542–545.
- Epperson BK, McRae BH, Scribner K *et al.* (2010) Utility of computer simulations in landscape genetics. *Molecular ecology*, **19**, 3549–3564.
- Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, **26**, 2064–2065.
- Excoffier L, Foll M (2011) fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, **27**, 1332–1334.
- Frichot E, François O (2015) LEA : An R package for landscape and ecological association studies. *Methods in ecology and evolution / British Ecological Society*, **6**, 925–929.
- Garza JC, Williamson EG (2001) Detection of reduction in population size using data from microsatellite loci. *Molecular ecology*, **10**, 305–318.
- Girod C, Vitalis R, Leblois R, Freville H (2011) Inferring Population Decline and Expansion From Microsatellite Data: A Simulation-Based Evaluation of the Msvar Method. *Genetics*, **188**, 165–179.
- Gruber B, Adamack AT (2015) landgenreport: a new r function to simplify landscape genetic analysis using resistance surface layers. *Molecular ecology resources*, **15**, 1172–1178.
- Haddock SHD, Dunn CW (2011) *Practical Computing for Biologists*. Sinauer Associates Incorporated.
- Hedrick PW (1995) Gene Flow and Genetic Restoration: The Florida Panther as a Case Study. *Conservation biology: the journal of the Society for Conservation Biology*, **9**, 996–1007.
- Hoban S (2014) An overview of the utility of population simulation software in molecular ecology. *Molecular ecology*, **23**, 2383–2401.
- Hoban S, Bertorelle G, Gaggiotti OE (2011) Computer simulations: tools for population and

- evolutionary genetics. *Nature reviews. Genetics*, **13**, 110–122.
- Hoban SM, Gaggiotti OE, Bertorelle G (2013a) The number of markers and samples needed for detecting bottlenecks under realistic scenarios, with and without recovery: a simulation-based study. *Molecular ecology*, **22**, 3444–3450.
- Hoban S, Sean H, Oscar G, Giorgio B, ConGRESS Consortium (2013b) Sample Planning Optimization Tool for conservation and population Genetics (SPOTG): a software for choosing the appropriate number of markers and samples. *Methods in ecology and evolution / British Ecological Society*, **4**, 299–303.
- Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.
- Jombart T, Cori A, Didelot X *et al.* (2014) Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS computational biology*, **10**, e1003457.
- Liu Y, Athanasiadis G, Weale ME (2008) A survey of genetic simulation software for population and epidemiological studies. *Human genomics*, **3**, 79–86.
- Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular ecology*, **24**, 1031–1046.
- Manel S, Stéphanie M, Schwartz MK, Gordon L, Pierre T (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in ecology & evolution*, **18**, 189–197.
- Marino IAM, Benazzo A, Agostini C *et al.* (2013) Evidence for past and present hybridization in three Antarctic icefish species provides new perspectives on an evolutionary radiation. *Molecular ecology*, **22**, 5148–5161.
- Münkemüller T, Tamara M, Sébastien L *et al.* (2012) How to measure and test phylogenetic signal. *Methods in ecology and evolution / British Ecological Society*, **3**, 743–756.
- Oyler-McCance SJ, Valdez EW, O'Shea TJ, Fike JA (2013) Genetic characterization of the Pacific sheath-tailed bat (*Emballonura semicaudata rotensis*) using mitochondrial DNA sequence data. *Journal of mammalogy*, **94**, 1030–1036.
- Peng B, Chen H-S, Mechanic LE *et al.* (2013) Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators. *Bioinformatics*, **29**, 1101–1102.
- Ryman N, Nils R, Stefan P (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular ecology notes*, **6**, 600–602.
- Segelbacher G, Gernot S, Cushman SA *et al.* (2010) Applications of landscape genetics in conservation biology: concepts and challenges. *Conservation genetics*, **11**, 375–385.
- Staab PR, Metzler D (2016) Coala: an R framework for coalescent simulation. *Bioinformatics*.
- Strand AE (2002) metasim 1.0: an individual-based environment for simulating population genetics of complex population dynamics. *Molecular ecology notes*, **2**, 373–376.
- Tallmon DA, Koyuk A, Luikart G, Beaumont MA (2008) COMPUTER PROGRAMS: onesamp: a program to estimate effective population size using approximate Bayesian computation. *Molecular ecology resources*, **8**, 299–301.