MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2017) **17**, 101–109 doi: 10.1111/1755-0998.12607

SPECIAL ISSUE: POPULATION GENOMICS WITH R

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

CHRISTIAN M. PAROBEK,* FREDERICK I. ARCHER,† MICHELLE E. DEPRENGER-LEVIN,‡ SEAN M. HOBAN,§ LIBBY LIGGINS¶ and ALLAN E. STRAND**

*Curriculum in Genetics and Molecular Biology, University of North Carolina, 135 Dauer Drive, 3206 Michael Hooker Research Center, Chapel Hill, NC 27599, USA, †Southwest Fisheries Science Center, 8901 La Jolla Shores Drive, La Jolla, CA 92037, USA, ‡Denver Botanic Gardens, 909 York Street, Denver, CO 80206, USA, §Morton Arboretum, 4100 Illinois Route 53, Lisle, IL 60532, USA, ¶Institute of Natural and Mathematical Sciences, Massey University, Auckland 0745, New Zealand, **College of Charleston, 66 George Street, Charleston, SC 29424, USA*

Abstract

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 geneticists and ecologists is hindered by difficulties in understanding these softwares' 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 genetic summary 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 genetic summary 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 role in molecular ecology investigations.** SKELESIM **can also serve as a teaching tool: demonstrating the outcomes of stochastic population genetic processes; teaching general concepts of simulations; and providing an introduction to the R environment with a user-friendly graphical user interface (using shiny).**

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

Received 31 May 2016; revision received 11 September 2016; accepted 26 September 2016

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. 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); testing the statistical performance of population genetic inference methods under different demographic scenarios (Girod *et al.* 2011; Hoban *et al.* 2013a); and evaluating spatial sampling strategies to infer the generating landscape process for spatial genetic patterns (Oyler-McCance *et al.* 2013; Lotterhos & Whitlock 2015). To generalize, simulations are used to create many in silico genetic data sets 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 data sets 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)

Correspondence Christian M. Parobek; E-mail: cmp@unc.edu

and to learn about the fundamentals of population genetics or decide on an appropriate sampling strategy, study method or management approach.

Dozens of software packages that implement simulations of demography, ecology, genetics, spatial processes, behaviour, adaptation, interspecies interactions and more now exist (Hoban *et al.* 2011; Peng *et al.* 2013). These software packages vary in complexity (Carvajal-Rodriguez 2008; Hoban *et al.* 2011), providing a wide range of options that 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 data sets, analyse 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 demographics, and POPPLANNER (Ewing & Hermisson 2010; Ewing *et al.* 2015) is a graphical tool which can be used to construct MS (Hudson 2002) and MSMS (Ewing & Hermisson 2010) command lines that model various scenarios. The downside of these software, however, is that they are specific to only one simulator, and they only construct the simulations and do not organize or analyse the datafiles. An prime example of an end-to-end solution is LandGenReport (Gruber & Adamack 2015), a comprehensive R package that 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). Other examples of user-friendly genetic simulation software 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). While these software help users analyse results, each is designed for a specific use of simulations, restricting users to certain scenarios and summary statistics. In contrast, COALA (Staab & Metzler 2016) is a recent R package that can wrap several coalescent simulators, standardizing the input and output files across programs and offering calculation of summary statistics. Such software that enables users to apply a single, convenient framework across different simulators would be very useful. 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, for different genetic markers (e.g. sequence, microsatellites, single nucleotide polymorphisms) and for a variety of current and potentially future genetic analyses.

Here, we provide an overview of 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 or knowledge base.

Materials and methods

Software features

SKELESIM is 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 genetic summary 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, analysed and summarized.

SKELESIM has two desirable features that do not exist in current simulation software. 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 (Fig. 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.

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, GGPLOT2 and RMETASIM (Jombart *et al.* 2016; Strand 2002; Jombart 2008; Wickham 2009; Paradis 2010; Archer *et al.* 2016). FASTSIMCOAL2 is a separate executable which must be manually installed (Excoffier & Foll 2011; Excoffier *et al.* 2013). Links and instructions are provided in the 'Installing' vignette in SKELESIM.

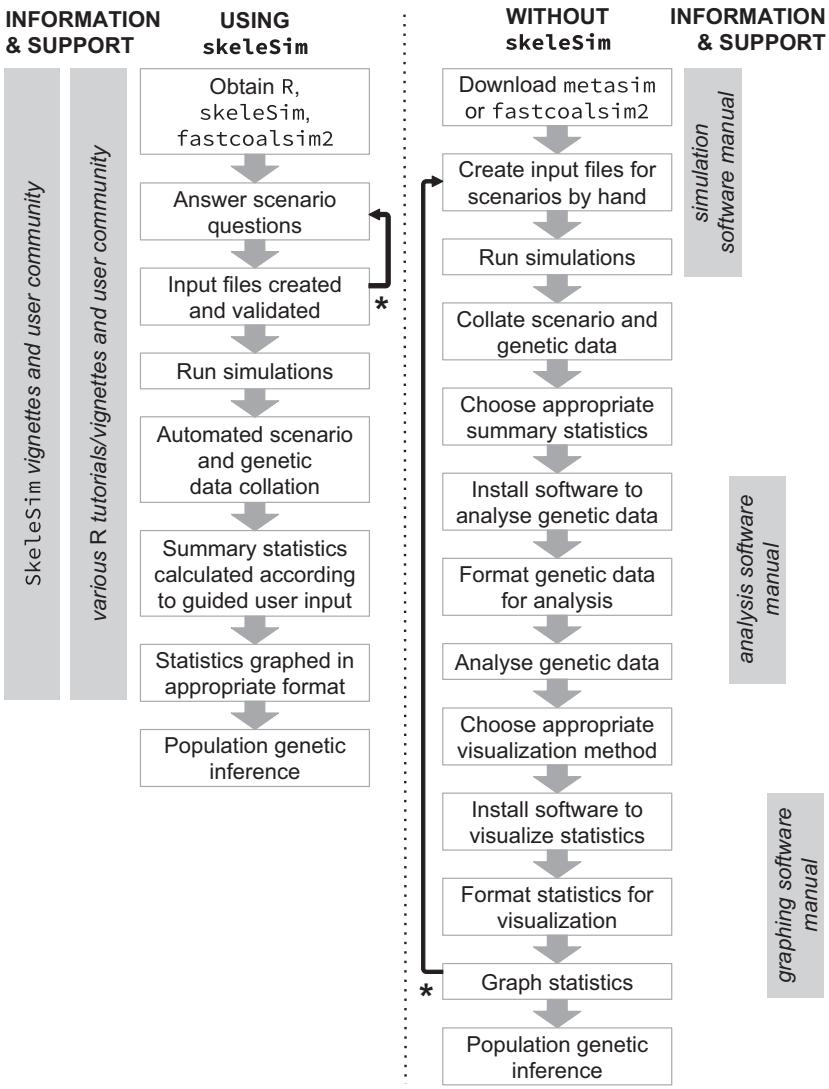

Fig. 1 An outline of the steps required to undertake a population genetic simulation study, and the sources of information and support available to the user when using SKELESIM versus the standalone METASIM or FASTSIMCOAL software. Using the R platform, SKELESIM provides centralized support for the user 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.

Currently implemented engines

SKELESIM can interface with simulation engines written either as native R packages (e.g. RMETASIM), or precompiled system executables (e.g. 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 *et al.* 2013)]. 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 software 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. Although both software 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 analyse complex scenarios with a reproducible pipeline. Thus, SKELESIM caters to the wide range of skill levels found within the molecular ecology community and may facilitate skill development at any stage 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 timescales are not needed. An additional difference is computational speed: coalescent simulators can

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 and 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

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 follow genotype frequencies rather than individuals. The distinctions among simulation categories are discussed in detail elsewhere (Carvajal-Rodriguez 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 'Simulations' provides an overview of the steps and describes the processes, labelling and construction of files that occur behind the graphical interface. The interface itself is organized in 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 genetic summary 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). Once the parameters of each scenario are saved, 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 to quickly visualize the genetic summary statistics for each simulated scenario and compare results among scenarios. The last tab of the interface is 'Current ssClass'. This tab allows the user to visualize 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 object-oriented coding conventions.

Architecture

All parameters of each scenario and results (including full output of all replicates and all genetic summary 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 reloaded in the shiny interface, or in any R environment (e.g. 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 genetic summary statistics within SKELESIM to describe simulation outputs. To help guide the selection of appropriate summary statistics, 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 customized 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*. 2016) and ADEGENET (Jombart 2008) that offer interoperability and complementary analysis options for population geneticists. A full list of genetic summary statistics available in the current version of SKELESIM is described in Table S1 (Supporting information).

Example use: forecasting case study

Example case studies for the use of SKELESIM are provided as vignettes that are downloaded with the package. This '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 either 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 (Fig. 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 Scenario 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 [Fig. 3; see Table S1 (Supporting information) 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 additional 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 simulation process, from setting up the models, to documenting pipelines, to obtaining organized results. It is often difficult for a user to know which parameters are required, whether their code or input files will run successfully and how to interpret error messages. While some

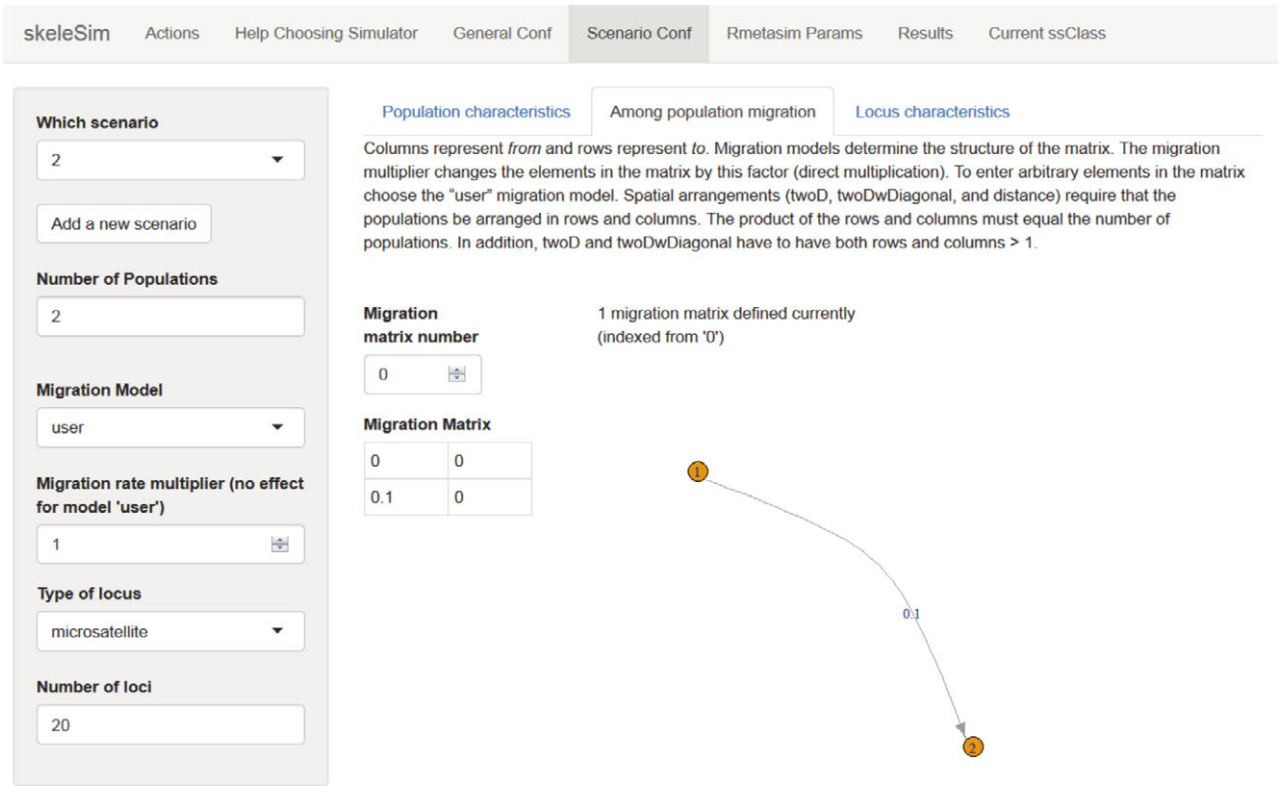

Fig. 2 The 'Scenario Conf' tab of the SKELESIM user interface. In this tab, the simulation scenarios are defined by the user. Migration rate and directionality are defined by the user in the matrix, and a matching population graph is automatically populated in the interface. This population graph corresponds to Scenario 2 of the 'Forecasting' example (see main text and SKELESIM vignettes). [Colour figure can be viewed at wileyonlinelibrary.com].

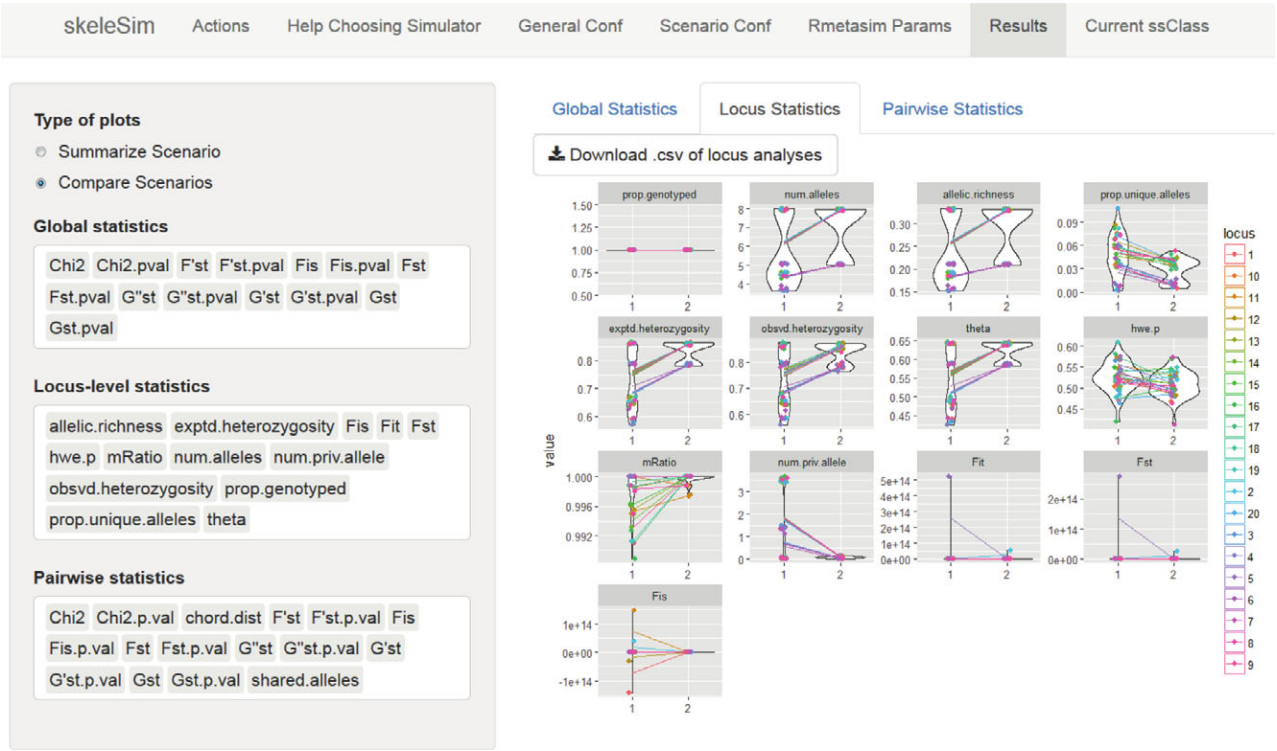

Fig. 3 The 'Results' tab of the SKELESIM user interface. Following the completion of simulations, users can upload and visualize the results for the genetic summary statistics they elected. In this 'Forecasting' example (see main text and SKELESIM vignettes), the results for the 'Locus Statistics' from 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. [Colour figure can be viewed at wileyonlinelibrary.com].

simulation software incorporate built-in analyses [e.g. METASIM; CDPOP (Landguth & Cushman 2010)], and some have visually accessible interfaces [e.g. PEDAGOG (Coombs *et al.* 2010)], 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 into other software, such as R, where one can use statistical and graphical approaches to draw inference from the simulations.

We have designed SKELESIM to fill this gap and provide a resource for both teaching and research. As a teaching tool, the SKELESIM GUI lends a gentle introduction to the R environment. It also provides a framework for learners to understand and compare the most common genetic markers (currently microsatellite, sequence and polymorphism data). As a resource for researchers, SKELESIM caters to the growing number of molecular ecologists who use R. Executing simulations within the familiar R environment will enable easy cross-application of coding skills and use of learning resources and fora already familiar to this community (Fig. 1). Furthermore, by 'wrapping' existing software in a common interface with streamlined terminology and inference output, population geneticists and molecular ecologists will be able to confidently switch between simulators. As the field of molecular ecology expands, SKELESIM fulfils a need of making software tools and analyses accessible to a wide audience.

SKELESIM will lessen 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 better understood and 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 rigour of studies in molecular ecology (Epperson *et al.* 2010; Balkenhol & Landguth 2011; Andrew *et al.* 2013). SKELESIM aims to remove the 'black box' perception associated with many

population genetic and simulation software. SKELESIM 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 and completely transparent. We 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 (Strand & Niehaus 2007), CDPOP and NEMO (Guillaume & Rougemont 2006), and coalescent software like MS. External developers can also add additional analyses to SKELESIM. 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 the current version of SKELESIM is natural selection. While future versions of SKELESIM may wrap simulators that include natural selection, this release serves as a proof-of-concept package that can wrap different simulators while facilitating the entire simulation process from parameter selection to result visualization. Thus, the current version of SKELESIM caters to the most common and tangible uses of population genetic simulations—generating '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 for identifying 'outlier' loci that do not fall within such distributions.

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, analysing 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. 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. Users are encouraged to fork the SKELESIM code from GITHUB and suggest or contribute to updates, new analyses and new simulators.

Acknowledgements

This project 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.

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 BB (2016) stratag: An r package for manipulating, summarizing and analysing population genetic data. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12559. [Epub ahead of print].

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 (2008) Simulation of genomes: a review. *Current Genomics*, **9**, 155–159.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2015) shiny: Web application framework for R [Software]. URL http://CRAN.R-project.org/package=shiny (R package version 0.12.2).

Coombs JA, Letcher BH, Nislow KH (2010) PEDAGOG: software for simulating eco-evolutionary population dynamics. *Molecular Ecology Resources*, **10**, 558–563.

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.

Ewing GB, Reiff PA, Jensen JD (2015) PopPlanner: visually constructing demographic models for simulation. In: *Frontiers in Genetics*, Vol. **6**. Frontiers editorial office, Lausanne.

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

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**, e1003905.

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.

Guillaume F, Rougemont J (2006) NEMO: an evolutionary and population genetics programming framework. *Bioinformatics*, **22**, 2556–2557.

Haddock SHD, Dunn CW (2011) *Practical Computing for Biologists*. Sinauer Associates Incorporated, Sunderland, Massachusetts.

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, Oscar G, Giorgio B (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*, **4**, 299–303.

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

Jombart T (2008) ADEGENET: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Jombart T, Archer F, Schliep K *et al*. (2016) apex: phylogenetics with multiple genes. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12567. [Epub ahead of print].

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.

Landguth EL, Cushman SA (2010) CDPOP: a spatially explicit cost distance population genetics program. *Molecular Ecology Resources*, **10**, 156–161.

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, Schwartz MK, Luikart G, Taberlet P (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.

Paradis E (2010) PEGAS: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**, 419–420.

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*, **32**, 1903–1904.

Strand AE (2002) METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics. *Molecular Ecology Notes*, **2**, 373–376.

Strand AE, Niehaus JM (2007) KERNELPOP, a spatially explicit population genetic simulation engine. *Molecular Ecology Notes*, **7**, 969–973.

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.

Wickham H (2009) GGPLOT2: *Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

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

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 in markdown form at (https://github.com/christianparobek/skeleSim/blob/master/vignettes)
- Stable version of SKELESIM and vignettes will be available from CRAN

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Overview of the genetic summary statistics available in the current version of SKELESIM

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

Parobek, CM

2017-01