Genetics and Genomics

Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
May 23, 2023 (this version)

Reviewed preprint version 1
March 3, 2023

Sent for peer review
December 21, 2022

Posted to preprint server
October 31, 2022

M. Elise Lauterbur, Maria Izabel A. Cavassim, Ariella L. Gladstein, Graham Gower, Nathaniel S. Pope, Georgia Tsambos, Jeff Adrion, Saurabh Belsare, Arjun Biddanda, Victoria Caudill, Jean Cury, Ignacio Echevarria, Benjamin C. Haller, Ahmed R. Hasan, Xin Huang, Leonardo Nicola Martin Iasi, Ekaterina Noskova, Jana Obšteter, Vitor Antonio Corrêa Pavinato, Alice Pearson, David Peede, Manolo F. Perez, Murillo F. Rodrigues, Chris C. R. Smith, Jeffrey P. Spence, Anastasia Teterina, Silas Tittes, Per Unneberg, Juan Manuel Vazquez, Ryan K. Waples, Anthony Wilder Wohns, Yan Wong, Franz Baumdicker, Reed A. Cartwright, Gregor Gorjanc, Ryan N. Gutenkunst, Jerome Kelleher, Andrew D. Kern, Aaron P. Ragsdale, Peter L. Ralph, Daniel R. Schrider, Ilan Gronau

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85719, USA • Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles CA, USA • Embark Veterinary, Inc., Boston MA 02111, USA • Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Denmark • Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA • School of Mathematics and Statistics, University of Melbourne, Australia • AncestryDNA, San Francisco CA 94107, USA • 54Gene, Inc., Washington DC 20005, USA • Université Paris-Saclay, CNRS, INRIA, Laboratoire Interdisciplinaire des Sciences du Numérique, UMR 9015 Orsay, France • School of Life Sciences, University of Glasgow, Glasgow, UK • Department of Computational Biology, Cornell University, Ithaca NY, USA • Department of Cell and Systems Biology, University of Toronto, Toronto ON, Canada • Department of Biology, University of Toronto Mississauga, Mississauga ON, Canada • Department of Evolutionary Anthropology, University of Vienna, Vienna, Austria • Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Vienna, Austria • Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany • Computer Technologies Laboratory, ITMO University, St Petersburg, Russia • Agricultural Institute of Slovenia, Department of Animal Science, Ljubljana, Slovenia • Entomology Department, The Ohio State University, Wooster OH, USA • Department of Genetics, University of Cambridge, Cambridge, UK • Department of Zoology, University of Cambridge, Cambridge, UK • Department of Ecology, Evolution, and Organismal Biology, Brown University, Providence RI, USA • Center for Computational Molecular Biology, Brown University, Providence RI, USA • Department of Genetics and Evolution, Federal University of Sao Carlos, Sao Carlos 13565905, Brazil • Department of Genetics, Stanford University School of Medicine, Stanford CA 94305, USA • Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Husargatan 3, SE-752 37 Uppsala, Sweden • Department of Integrative Biology, University of California, Berkeley, Berkeley CA, USA • Department of Biostatistics, University of Washington, Seattle WA, USA • Broad Institute of MIT and Harvard, Cambridge MA 02142, USA • Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, UK • Cluster of Excellence - Controlling Microbes to Fight Infections, Eberhard Karls Universität Tübingen, Tübingen, Baden-Württemberg, Germany • School of Life Sciences and The Biodesign Institute, Arizona State University, Tempe AZ, USA • The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh EH25 9RG, UK • Department of Molecular and Cellular Biology, University of Arizona, Tucson AZ 85721, USA • Department of Integrative Biology, University of Wisconsin-Madison, Madison WI, USA • Department of Mathematics, University of Oregon, Eugene OR 97402, USA • Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill NC 27599, USA • Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel

Abstract

Simulation is a key tool in population genetics for both methods development and empirical research, but producing simulations that recapitulate the main features of genomic data sets remains a major obstacle. Today, more realistic simulations are possible thanks to large increases in the quantity and quality of available genetic data, and to the sophistication of inference and simulation software. However, implementing these simulations still requires substantial time and specialized knowledge. These challenges are especially pronounced for simulating genomes for species that are not well-studied, since it is not always clear what information is required to produce simulations with a level of realism sufficient to confidently answer a given question. The community-developed framework stdpopsim seeks to lower this barrier by facilitating the simulation of complex population genetic models using up-to-date information. The initial version of stdpopsim focused on establishing this framework using six well-characterized model species (Adrion et al., 2020). Here, we report on major improvements made in the new release of stdpopsim (version 0.2), which includes a significant expansion of the species catalog and substantial additions to simulation capabilities. Features added to improve the realism of the simulated genomes include non-crossover recombination and provision of species-specific genomic annotations. Through community-driven efforts, we expanded the number of species in the catalog more than three-fold and broadened coverage across the tree of life. During the process of expanding the catalog, we have identified common sticking points and developed best practices for setting up genome-scale simulations. We describe the input data required for generating a realistic simulation, suggest good practices for obtaining the relevant information from the literature, and discuss common pitfalls and major considerations. These improvements to stdpopsim aim to further promote the use of realistic whole-genome population genetic simulations, especially in non-model organisms, making them available, transparent, and accessible to everyone.

eLife assessment

This **important** paper reports recent improvements and extensions to stdpopsim, a community-driven resource that is built on top of powerful software for performing simulations of population genomic data and provides a catalog of species with curated genomic parameters and demographic models. In addition to describing the new features and species in stdpopsim, the authors provide a set of practical guidelines for implementing realistic simulations. Overall, this **convincing** manuscript serves as an excellent overview of the utility, challenges, common pitfalls, and best practices of population genomic simulations. It will be of broad interest to population, evolutionary, and ecological geneticists studying humans, model organisms, or non-model organisms.

Introduction

Population genetics allows us to answer questions across scales from deep evolutionary time to ongoing ecological dynamics, and dramatic reductions in sequencing costs enable the generation of unprecedented amounts of genomic data that can be used to address these questions (Ellegren, 2014). Ongoing efforts to systematically sequence life on Earth by initiatives such as the Earth

Biogenome (Lewin et al., 2022) and its affiliated project networks, such as Vertebrate Genomes (Rhie et al., 2021), 10,000 Plants (Cheng et al., 2018) and others (Darwin Tree of Life Project Consortium, 2022), are providing the backbone for enormous increases in the amount of population-level genomic data available for model and non-model species. These data are being used, among other things, in inference of population history and demographic parameters (Beichman et al., 2018), studying adaptive introgression (Gower et al., 2021), providing null expectations for selection scans (e.g. Hsieh et al., 2021), and understanding the implications of deleterious variation in populations of conservation concern (e.g. Robinson et al., 2023). While many of the methods that address these questions were initially developed for a few key model systems such as humans and *Drosophila*, more recent efforts are generalizing these methods to include important factors not initially accounted for, such as inbreeding or selfing (Blischak et al., 2020), skewed offspring distributions (Montano, 2016), and intense artificial selection even for non-model organisms (MacLeod et al., 2013, 2014).

Simulations can be useful at all stages of this work—for planning studies, analyzing data, testing inference methods, and validating findings from empirical and theoretical research. For instance, simulations provide training data for inference methods based on machine learning (Schrider and Kern, 2018) and Approximate Bayesian Computation (Csilléry et al., 2010). They can also serve as baselines for further analyses: for example, simulations incorporating demographic history serve as null models when detecting selection (Hsieh et al., 2016) or seed downstream breeding program simulations (Gaynor et al., 2020). More recently, population genomic simulations have been used to help guide conservation decisions for threatened species (Teixeira and Huber, 2021; Kyriazis et al., 2022).

Increasing amounts of data and sophistication of inference methods have enabled researchers to ask ever more specific and precise questions. Consequently, simulations must incorporate more and more elements of biological realism. Important elements include genomic features such as mutation and recombination rates that strongly affect genetic variation and haplotype structure (Nachman, 2002). The inclusion of these genomic features is particularly important when linked selection is acting upon the patterns of genomic diversity being studied (Cutter and Payseur, 2013). Furthermore, the demographic history of a species—encompassing population sizes and distributions, divergences, and gene flow—can dramatically affect patterns of genomic variation (Teshima et al., 2006). Thus species-specific estimates of these and other ecological and evolutionary parameters (such as those governing the process of natural selection) are important when generating realistic simulations. This presents challenges, especially to new researchers, as it takes a great deal of specialized knowledge not only to code the simulations themselves but also to find and choose appropriate estimates of the parameters underlying the simulation model.

The recently developed community resource stdpopsim provides easy access to detailed population genomic simulations (Adrion et al., 2020). It lowers the technical barriers to performing these simulations and reduces the possibility of erroneous implementation of simulations for species with published demographic models. The initial release of stdpopsim was restricted to only six well-characterized model species, such as *Drosophila melanogaster* and *Homo sapiens*, but feedback we received from the community identified a widespread desire to simulate a broader range of non-model species, and ideally to incorporate these into the stdpopsim catalog for future use. This feedback, and subsequent efforts to expand the catalog, also uncovered a vital need to better understand when it is practical to create a realistic simulation of a species of interest, and indeed what “realistic” means in this context.

This paper reports on the updates made in the current release of stdpopsim (version 0.2), and is also intended as a resource for any researcher who wishes to develop chromosome-scale simulations for their own species of interest. We start by describing the central idea behind the standardized simulation framework of stdpopsim, and then outline the main updates made to the stdpopsim catalog and simulation framework in the past two years. We then provide guidelines

for generating population genomic simulations, either for the purpose of using them in one specific study, or with the intent of making the simulations available for future work by adding the appropriate models to stdpopsim. Among other considerations, we discuss when a chromosome-scale simulation is more useful than simulations based on either individual loci or generic loci. We specify the required input data, mention common pitfalls in choosing appropriate parameters, and suggest courses of action for species that are missing estimates of some necessary inputs. We conclude with examples from two species recently added to stdpopsim, which demonstrate some of the main considerations involved in the process of designing realistic chromosome-scale simulations. While the guidelines provided in this paper are intended for any researcher interested in implementing a population genomic simulation using any software, we highlight the ways in which the stdpopsim framework eases the burden involved in this process and facilitates reproducible research.

The utility of stdpopsim for chromosome-scale simulations

We begin by providing a brief overview of the importance of chromosome-scale simulations and the main rationale behind stdpopsim; see Adrion et al. (2020) for more on the topic. The main objective of population genomic simulations is to recreate patterns of sequence variation along the genome under the inferred evolutionary history of a given species. To achieve this, stdpopsim is built on top of the msprime (Kelleher et al., 2016; Nelson et al., 2020; Baumdicker et al., 2021) and SLiM (Haller and Messer, 2019) simulation engines, which are capable of producing fairly realistic patterns of sequence variation if provided with accurate descriptions of the genome architecture and evolutionary history of the simulated species. The required parameters include the number of chromosomes and their lengths, mutation and recombination rates, the demographic history of the simulated population, and, potentially, the landscape of natural selection along the genome. A key challenge when setting up a population genomic simulation is to obtain estimates of all of these quantities from the literature and then correctly implement them in an appropriate simulation engine. Detailed estimates of all of these quantities are increasingly available due to the growing availability of population genomic data coupled with methodological advances. Incorporating this data into a population genomic simulation often involves integrating this data between different literature sources, which can require specialized knowledge of population genetics theory. Thus, the process of coding a realistic simulation can be quite time-consuming and often error-prone.

The main objective of stdpopsim is to streamline this process, and to make it more robust and more reproducible. Contributors collect parameter values for their species of interest from the literature, and then specify these parameters in a template file for the new model. This model then undergoes a peerreview process, which involves another researcher independently recreating the model based on the provided documentation. Automated scripts then execute to compare the two models; if discrepancies are found in this process, they are resolved by discussion between the contributor and reviewer, and if necessary with input of additional members of the community. This quality-control process quite often finds subtle bugs (e.g., as in Ragsdale et al., 2020) or highlights parts of the model that are ambiguously defined by the literature sources. This increases the reliability and reproducibility of the resulting simulations in any downstream analysis.

Another important goal of stdpopsim is to promote and facilitate chromosome-scale simulations, as opposed to the common practice of simulating many short segments (see, e.g., Harris and Nielsen, 2016). Simulation of long sequences, on the order of 10^7 bases, has until recently been computationally prohibitive, but this has changed with the development of modern simulation engines such as msprime and SLiM. Generating chromosome-scale simulations has several key

benefits. First, the organization of genes on chromosomes is a key feature of a species' genome that is ignored in many traditional population genomic simulations (see Schrider (2020) for one exception). Second, modeling physical linkage allows simulations to capture important correlations between genetic variants on a chromosome. These correlations reduce variance relative to separate and independent simulations of equivalent genetic material. This has a particularly striking effect in long stretches with a low recombination rate, as observed for instance on the long arm of human chromosome 22 (Dawson et al., 2002). In bacteria, a similar effect occurs due to genome-wide linkage that is broken only by horizontal transfer of short segments (Didelot and Maiden, 2010). When conducting simulations with natural selection, linkage has an even stronger effect. Selection acting on a small number of sites can indirectly influence levels and patterns of genetic variation at linked neutral sites, which has been shown to have a widespread effect on patterns of genomic variation in a myriad of species (e.g., McVicker et al., 2009; Charlesworth, 2012). In addition, the lengths of chromosome-scale shared haplotypes within and between populations provides valuable information on their demographic history. Demographic inference methods that use such information, such as MSMC (Schiffels and Wang, 2020) and IBDNe (Browning and Browning, 2015), perform best on long genomic segments with realistic recombination rates. Chromosome-scale simulations are clearly required to test (or train) such methods, or to conduct power analyses when designing empirical studies that use them. With stdpopsim, such simulations are available with just a single call to a command-line script or with execution of a handful of lines of Python code.

Additions to stdpopsim

When first published, the stdpopsim catalog included six species: *Homo sapiens, Pongo abelii, Canis familiaris, Drosophila melanogaster*, *Arabidopsis thaliana*, and *Escherichia coli* (**Figure 1**). One way the catalog has expanded is through the introduction of additional demographic models for *Homo sapiens*, *Pongo abelii*, *Drosophila melanogaster*, and *Arabidopsis thaliana*, enabling a wider variety of simulations for these well-studied species. However, the initial collection of six species represents only a small slice of the tree of life. This is a concern not only because there is a large community of researchers studying other organisms, but also because methods developed for application to model species (such as humans) may not perform well when applied to other species with very different biology. Adding species to the stdpopsim catalog will allow developers to easily test their methods across a wider variety of organisms.

We thus made a concerted effort to recruit members of the population and evolutionary genetics community to add their species of interest to the stdpopsim catalog. This effort involved a series of workshops to introduce potential contributors to stdpopsim, followed by a "Growing the Zoo" hackathon organized alongside the 2021 ProbGen conference. The seven initial workshops allowed us to reach a broad community of more than 150 researchers, many of whom expressed interest in adding non-model species to stdpopsim. The hackathon was then structured based on feedback from these participants. One month before the hackathon, we organized a final workshop to prepare interested participants, by introducing them to the process of developing a new species model and adding it to the stdpopsim code base. Roughly 20 scientists participated in the hackathon (most of whom are included as authors on this paper), which resulted in the addition of 15 species to the stdpopsim catalog (**Figure 1**). The catalog now includes a teleost fish (*Gasterosteus aculeatus*), a bird (*Anas platyrhynchos*), a reptile (*Anolis carolinensis*), a livestock species (*Bos taurus*), six insects including two vectors of human disease (*Aedes aegypti* and *Anopheles gambiae*), a nematode (*Caenorhabditis elegans*), two flowering plants including a crop (*Helianthus annuus*), an algae (*Chlamydomonas reinhardtii*), two bacteria, four primates, and a common mammalian associate of humans (*Canis familiaris*). Not all of these have recombination maps or demographic models (see **Figure 1**), but this lays a framework for future contributions.

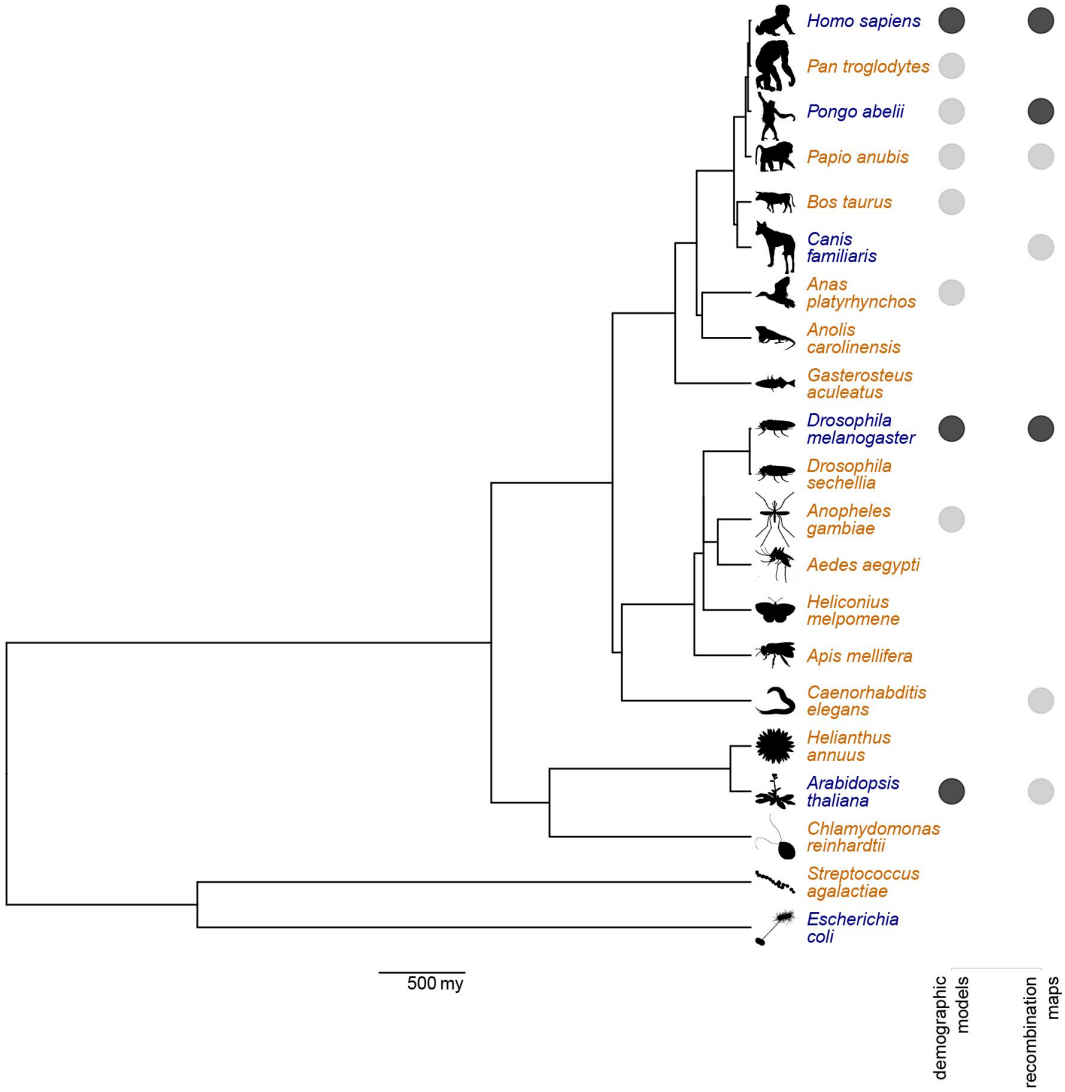

Figure 1

Phylogenetic tree of species available in the stdpopsim catalog, including the six species we published in the original release (Adrion et al., 2020, in blue), and 15 species that have since been added (in orange). Solid circles indicate species that have one (light grey) or more (dark grey) demographic models and recombination maps. Branch lengths were derived from the divergence times provided by TimeTree5 (Kumar et al., 2022). The horizontal bar below the tree indicates 500 million years (my).

Expanding the species catalog required adding several capabilities to the simulation framework of stdpopsim. Some features were added by upgrading the neutral simulation engine, msprime, from version 0.7.4 to version 1.0 (Baumdicker et al., 2021). Among other features, this upgrade includes a discrete-site model of mutation, which enables simulating sites with multiple mutations and possibly more than two alleles. Another key feature added to stdpopsim's simulation framework was the ability to model non-crossover recombination. In bacteria and archaea, genetic material can be exchanged through horizontal gene transfer, which can add new genetic material (e.g., via the transfer of plasmids) or replace homologous sequences through homologous recombination (Thomas and Nielsen, 2005; Didelot and Maiden, 2010; Gophna and Altman-Price, 2022). However, the initial version of stdpopsim used crossover recombination to stand in for these processes. Although we cannot currently simulate varying gene content (as would be required to simulate the addition of new genetic material by horizontal gene transfer), the msprime and SLiM simulation engines now allow gene conversion, which has the same effect as non-crossover homologous recombination. Following Cury et al. (2022), we use this to include non-crossover homologous recombination in bacterial and archaeal species. This is done in stdpopsim by setting a flag in the species model to indicate that recombination should be modeled without crossovers, and specifying an average tract length of exchanged genetic material. For example, the model for *Escherichia coli* has been updated in the stdpopsim catalog to use non-crossover recombination at an average rate of 8.9×10^{-11} recombination events per base per generation, with an average tract length of 542 bases (Wielgoss et al., 2011; Didelot et al., 2012). Note that this rate (8.9×10^{-11}) corresponds to the rate of initiation of a recombined tract.

Recombination without crossover is also prevalent in sexually reproducing species, where it is termed *gene conversion*. Gene conversion affects shorter segments than crossover recombination and creates distinct patterns of genetic diversity along the genome (Korunes and Noor, 2017). Indeed, gene conversion rates in some species are estimated to occur at similar or even higher rates than crossover recombination (Gay et al., 2007; Comeron et al., 2012; Wijnker et al., 2013). To accommodate this in stdpopsim simulations, one needs to specify the fraction of recombinations that occur due to gene conversion (i.e., without crossover), and the average tract length. For example, the model for *Drosophila melanogaster* has been updated in the stdpopsim catalog to have a fraction of gene conversions of 0.83 (in all chromosomes that undergo recombination) and an average tract length of 518 bases (Comeron et al., 2012). This update does not affect the rate of crossover recombination, but it adds gene conversion events at a ratio of 83:17 relative to crossover recombination events. We note that since non-crossover recombination incurs a high computation load in simulation, it is turned off by default in stdpopsim, and must be explicitly invoked by the simulation model. Note that ignoring gene conversion may result in a slightly skewed distribution of shared haplotypes between individuals (see **Table 1**).

Another important extension of stdpopsim allows augmenting a genome assembly with genome annotations, such as coding regions, promoters, and conserved elements. These annotations can be used to simulate selection at a subset of sites (such as the annotated coding regions) using parametric distributions of fitness effects. Standardized, easily accessible simulations that include the reality of pervasive linked selection in a species-specific manner has long been identified as a goal for evolutionary genetics (e.g., McVicker et al., 2009; Comeron, 2014). Thus, we expect this extension of stdpopsim to be transformative in the way simulations are carried out in population genetics. This significant new capability of the stdpopsim library will be detailed in a forthcoming publication, and is not the focus of this paper.

Missing parameter	Suggested action	Possible discrepancies
Mutation rate	Borrow from closest relative with a citable mutation rate	Number of polymorphic sites
Recombination rate	Borrow from closest relative with a citable recombination rate	Patterns of linkage disequilibrium
Gene conversion rate and tract length	Set rate to 0 or borrow from closest relative with a citable rate	Lengths of shared haplotypes across individuals
Demographic model	Set the effective population size (N_e) to a value that reflects the observed genetic diversity	Features of genetic diversity that are captured by the site frequency spectrum, such as the prevalence of low-frequency alleles

Table 1

Guidelines for dealing with missing parameters.

For each parameter, we provide a suggested course of action, and mention the main discrepancies between simulated data and real genomic data that could be caused by misspecification of that parameter.

Guidelines for implementing a population genomic simulation

The concentrated effort to add species to the stdpopsim catalog has led to a series of important insights about this process, which we summarize here as a set of guidelines for implementing realistic simulations for any species. Our intention is to provide general guidance that applies to any population genomic simulation software, but we also mention specific requirements that apply to simulations done in stdpopsim.

Basic setup for chromosome-level simulations

Implementing a realistic population genomic simulation for a species of interest requires a detailed description of the organism's demography and mechanisms of genetic inheritance. While simulation software requires unforgivingly precise values, in practice we may only have rough guesses for most of the parameters describing these processes. In this section, we list the relevant parameters and provide guidelines for how to set them based on current knowledge.

1. **A chromosome-level genome assembly**, which consists of a list of chromosomes or scaffolds and their lengths. Having a good quality assembly with complete chromosomes, or at least very long scaffolds, is necessary if chromosome-level population genomic simulations are to reflect the genomic architecture of the species. When expanding the stdpopsim catalog during the "Growing the Zoo" hackathon, we considered the possibility of adding species whose genome assemblies are composed of many relatively small contigs, unanchored to chromosome-level scaffolds. Although we had not previously put restrictions on which species might be added, we decided that we would only add species with chromosome-level assemblies. The main justification for this restriction is that species with less complete genome builds typically do not have good recombination maps and demographic models, making chromosome-level simulation much less useful in such species. Another issue is the storage burden and long load times involved in dealing with hundreds of contigs. Finally, each species requires validation of its code before it is added to the stdpopsim catalog, as well as long-term maintenance to keep it up-to-date with changes made to the stdpopsim framework. So, the benefit of including species with very partial genome builds in stdpopsim would be outweighed by the substantial extra burden on stdpopsim maintainers as well as downstream users of these models. Another reason to focus on species with chromosome-level assemblies is that we expect their numbers to dramatically increase in the near future due to numerous genome initiatives (Lewin et al., 2022; Rhie et al., 2021; Cheng et al., 2018) and the development of new long-read sequencing technologies and assembly pipelines (Chakraborty et al., 2016; Amarasinghe et al., 2020, 2021).
2. **An average mutation rate** for each chromosome (per generation per bp). This rate estimate can be based on sequence data from pedigrees, mutation accumulation studies, or comparative genomic analysis calibrated by fossil data (i.e., phylogenetic estimates). At present, stdpopsim simulates mutations at a constant rate under the Jukes–Cantor model of nucleotide mutations (Jukes and Cantor, 1969). However, we anticipate future development will provide support for more complex, heterogeneous mutational processes, as these are easily specified in both the SLiM and msprime simulation engines. Such progress will further improve the realism of simulated genomes, since mutation processes, including rates, are known to vary along the genome and through time (Benzer, 1961; Ellegren et al., 2003; Supek and Lehner, 2019).

3. **Recombination rates** (per generation per bp). Ideally, a population genomic simulation should make use of a chromosome-level recombination map, since the recombination rate is known to vary widely across chromosomes (Nachman, 2002), and this can strongly affect the patterns of linkage disequilibrium and shared haplotype lengths. When this information is not available, we suggest specifying an average recombination rate for each chromosome. At minimum, an average genomewide recombination rate needs to be specified, which is typically available for well-assembled genomes. For bacteria and archaea, which primarily experience non-crossover recombination, the average tract length should also be specified (see details in previous section). **Gene conversion (optional):** If one wishes to model gene conversion in eukaryotes, either together with crossover recombination or as a stand-alone process, then one should specify the fraction of recombinations done by gene conversion as well as the per chromosome average tract length.
4. **A demographic model** describing ancestral population sizes, split times, and migration rates. Selection of a reasonable demographic model is often crucial, since misspecification of the model can generate unrealistic patterns of genetic variation that will affect downstream analyses (e.g., Navascués and Emerson, 2009). A given species might have more than one demographic model, fit from different data or by different methods. Thus, when selecting a demographic model, one should examine the data sources and methods used to obtain it to ensure that they are relevant to the study at hand (see also **Limiations of simulated genomes** below). At a minimum, simulation requires a single estimate of effective population size. This estimate, which may correspond to some sort of historical average effective population size, should produce simulated data that matches the average observed genetic diversity in that species. Note, however, that this average effective population size cannot capture features of genetic variation that are caused by recent changes in population size and the presence of population structure (MacLeod et al., 2013; Eldon et al., 2015). For example, a recent population expansion will produce an excess of low-frequency alleles that no simulation of a constant-sized population will reproduce (Tennessen et al., 2012).
5. **An average generation time** for the species. This parameter is an important part of the species' natural history. This value does not directly affect the simulation, since stdpopsim uses either the Wright–Fisher model (in SLiM) or the Moran model (in msprime), both of which operate in time units of generations. Thus, the average generation time is only currently used to convert time units to years, which is useful when comparing among different demographic models.

These five categories of parameters are sufficient for generating simulations under neutral evolution. Such simulations are useful for a number of purposes, but they cannot be used to model the influence of natural selection on patterns of genetic variation. To achieve this, the simulator needs to know which regions along the genome are subject to selection, and the nature and strength of this selection. As mentioned above, the ability to simulate chromosomes with realistic models of selection is still under development, and will be finalized in the next release of stdpopsim. The development version of stdpopsim enables simulation with selection (using the SLiM engine) by specifying genome annotations and distributions of fitness effects, as specified below.

- 6. **Genome annotations**, specifying regions subject to selection (as, for example, a GFF3/GTF file). For instance, annotations can contain information on the location of coding regions, the position of specific genes, or conserved non-coding regions. Regions not covered by the annotation file are assumed to be evolving free from the effects of direct natural selection.
- 7. **Distributions of fitness effects** (DFEs) for each annotation. Each annotation is associated with a DFE describing the probability distribution of selection coefficients (deleterious, neutral, and beneficial) for mutations occurring in the region covered by the

annotation. DFEs can be inferred from population genomic data (reviewed in Eyre-Walker and Keightley, 2007), and are available for several species (e.g., Ma et al., 2013; Huber et al., 2018).

The current release of stdpopsim contains annotations and implemented DFE models for the three model species: *A. thaliana*, *D. melanogaster*, and *H. sapiens*. A forthcoming publication will provide details about how this is implemented in stdpopsim and examples of possible uses of this feature.

Extracting parameters from the literature

Simulations cannot of course precisely match reality, but in setting up simulations it is desirable to choose parameters that best reflect our current understanding of the evolutionary history of the species of interest. In practice a researcher may choose each parameter to match a fairly precise estimate or a wild guess, which may be obtained from a peer-reviewed publication or by word of mouth. However, values in stdpopsim are always chosen to match published estimates, so that the underlying data and methods are documented and can be validated. Because the process of converting information reported in the literature to parameters used by a simulation engine is quite error-prone, independent validation of the simulation code is crucial. We highly recommend following a quality-control procedure similar to the one used in stdpopsim, in which each species or model added to the catalog is independently recreated or thoroughly reviewed by a separate researcher.

Obtaining reliable and citable estimates for all model parameters is not a trivial task. Oftentimes, values for different parameters must be gleaned from multiple publications and combined. For example, it is not uncommon to find an estimate of a mutation rate in one paper, a recombination map in a separate paper, and a suitable demographic model in a third paper. Integrating information from different publications requires caution, since some of these parameter estimates are entangled in non-trivial ways. For instance, consider simulating a demographic model estimated in a specific paper that assumes a certain mutation rate. Naively using the demographic model, as published, with a new estimate of the mutation rate will lead to levels of genetic diversity that do not fit the genomic data. This is addressed in stdpopsim by allowing a demographic model to be simulated using a mutation rate that differs from the default rate specified for the species. See, for example, the model implemented for *Bos taurus*, which is described in the next section. This important feature does not necessarily fix all potential inconsistencies caused by assumptions made by the demographic inference method (such as assumptions about recombination rates). It is therefore recommended, when possible, to take the demographic model, mutation rates, and recombination rates from the same study, and to proceed carefully when mixing sources. An additional tricky source of inconsistency is coordinate drift between subsequent versions of genome assemblies. In stdpopsim, we follow the UCSC Genome Browser and use liftOver to convert the coordinates of recombination maps and genome annotations to the coordinates of the current genome assembly (Hinrichs et al., 2006).

Limitations of simulated genomes

Despite their great utility, simulated genomes cannot fully capture all aspects of genetic variation as observed in real data, with some aspects modeled better than others. As mentioned above, this will strongly depend on the demographic model used in simulation. Thus, it is important to consider potential limitations of different demographic models in reflecting observed genetic variation. First, a demographic model inferred from analysis of genomic data will likely depend on the samples that contributed the analyzed genomes. The inferred demographic model can only reflect the genealogical ancestry of these sampled individuals, and this will typically make up a small portion of the complete genealogical ancestry of the species. Thus, demographic models inferred from larger sets of samples from diverse ancestry backgrounds may potentially provide a more comprehensive depiction of genetic variation within a species. This is true if sufficiently realistic demographic models can be fit—models that account for the structure of populations

within a species. That said, the choice of samples used for inference will mostly influence recent changes in genetic variation. This is because the genealogy of even a single individual consists of numerous ancestors in each generation in the deep past, which is the premise of methods that infer ancestral population sizes from a single input genome (Li and Durbin, 2011).

The computational method used for inference also affects the way genetic variation is reflected by the demographic model, because different methods derive their inference from different features of genomic variation. Some methods make use of the site frequency spectrum at unlinked single sites (e.g., Gutenkunst et al., 2009; Excoffier et al., 2013; Liu and Fu, 2015), while other methods use haplotype structure (e.g., Li and Durbin, 2011; Schiffels and Wang, 2020; Browning and Browning, 2015). This, in turn, may influence the accuracy of different features in the inferred demography. For example, very recent demographic changes, such as recent admixture or bottlenecks, are difficult to infer from the site frequency spectrum, but are more easily inferred by examining shared long haplotypes (as demonstrated by the demographic model inferred for *Bos taurus* by MacLeod et al. (2013); see below). Several studies have compared different approaches to demographic inference (e.g., Harris and Nielsen, 2013; Beichman et al., 2017), but unfortunately, there is currently no succinct handbook that describes the relative strengths and weaknesses of different methods. Thus, assessing the potential limitations of a given demographic model currently requires some familiarity with the method used for its inference. In addition, all methods assume that the input sequences are neutrally evolving. This implies that technical choices, such as the specific genomic segments analyzed and various filters, may also influence the inferred model and its ability to model observed genetic variation. Thus, it is strongly advised to read the study that inferred the demographic model and understand potential limitations that stem from the selection of samples, methods, and filters.

We note that inclusion of a demographic model in the stdpopsim catalog does not involve any judgment as to which aspects of genetic variation it captures. Any model that is a faithful implementation of a published model inferred from genomic data can be added to the stdpopsim catalog. Thus, potential users of stdpopsim should use the implemented models with the appropriate caution, keeping in mind the limitations discussed above. We maintain a fairly detailed documentation page for the catalog (see **Data availability**), which contains a brief summary for each demographic model. This summary includes a graphical description of the model (such as the one shown for *Anopheles gambiae* in **Fig. 2B**), as well as a description of the data and method used for inference. Therefore, the documentation can provide guidance to potential users of stdpopsim in the process of selecting an appropriate demographic model for simulation. Finally, we hope that the standardized simulations implemented in stdpopsim will facilitate additional studies that examine the relative strengths and limitations of different approaches to demographic inference (and modeling genetic variation in general), and this will allow us to generate even more realistic simulations in the future.

Filling in the missing pieces

For many species it is difficult to obtain estimates of all necessary model parameters. **Table 1** provides suggestions for ways to deal with missing values of various model parameters. The table also mentions possible consequences of misspecification of each parameter.

In some cases, one may wish to generate simulations for a species with a partial genome build. Despite the focus of stdpopsim on species with chromosome-level assemblies (see discussion above), simulation is still potentially useful for species with less complete assemblies, with some important considerations to keep in mind. Longer contigs or scaffolds in these builds can be simulated separately and independently. This approach allows us to model genetic linkage within each contig, but linkage between different contigs that map to the same chromosome will not be captured by the simulation. This provides a reasonable approximation for many purposes, at least for genomic regions far from the contig edges. For shorter contigs, separate independent

A

Chromosome	Chromosome length	Recombination rate	Mutation rate
2L	49,364,325	1.30e−08	3.5e−09
2R	61,545,105	1.30e−08	3.5e−09
3L	41,963,435	1.30e−08	3.5e−09
3R	53,200,684	1.60e−08	3.5e−09
X	24,393,108	2.04e−08	3.5e−09
Mt	15,363	0.00e+00	3.5e−09

B

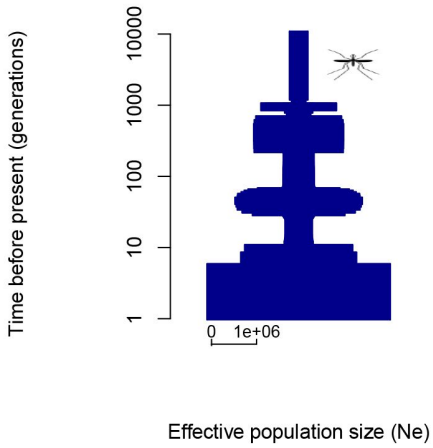

Figure 2

The species parameters and demographic model used for *Anopheles gambiae* in the stdpopsim catalog. (A) The parameters associated with the genome build and species, including chromosome lengths, average recombination rates (per base per generation), and average mutation rates (per base per generation). (B) A graphical depiction of the demographic model, which consists of a single population whose size changes throughout the past 11,260 generations in 67 time intervals (note the log scale). The width at each point depicts the effective population size (N_e), with the horizontal bar at the bottom indicating the scale for $N_e = 10^6$. This figure is adapted from the data on the stdpopsim catalog documentation page (see **Data availability**) and plotted with POPdemog (Zhou et al., 2018).

simulations will not be able to capture patterns of long-range linkage in a reasonably realistic way. Thus, a potentially viable option for shorter contigs is to combine them into longer pseudochromosomes, trying to mimic the species' expected chromosome lengths. Despite their somewhat artificial construction, these pseudo-chromosomes have the important benefit of capturing patterns of linkage similar to those observed in real genomic chromosomes. If, for example, the main purpose of the simulation is to examine the distribution of lengths of shared haplotypes between individuals, or study patterns of background selection, then it makes sense to simulate such pseudo-chromosomes. However, genetic correlations between different specific contigs lumped together in this way are obviously not accurate. So, if the main purpose of the simulation is to examine local patterns of genetic variation in loci of interest, then it may be more appropriate to simulate the relevant contigs separately (even if they are short), or to randomly sample several mappings of contigs to pseudo-chromosomes. For some purposes it makes sense to simulate a large number of unlinked sites (Gutenkunst et al., 2009; Excoffier et al., 2013), which can be generated without any sort of genome assembly. However, this approach would not have the benefits of chromosome-scale simulations. While some of the same considerations hold when simulating unlinked short sequences, a detailed discussion about such simulations goes beyond the scope of this paper. Ultimately, the recommended mode of simulation for a species with a partial genome assembly depends on the intended use of the simulated genomes.

Examples of added species

In this section, we provide examples of two species recently added to the stdpopsim catalog, *Anopheles gambiae* and *Bos taurus*, to demonstrate some of the key considerations of the process. In each example, we highlight in bold the model parameters set for each species.

Anopheles gambiae (mosquito)

Anopheles gambiae, the African malaria mosquito, is a non-model organism whose population history has direct implications for human health. Several large-scale studies in recent years have provided information about the population history of this species on which population genomic simulations can be based (e.g., Miles et al., 2017; Clarkson et al., 2020). The genome assembly structure used in the species model is from the AgamP4 **genome assembly** (Sharakhova et al., 2007), downloaded directly from Ensembl (Howe et al., 2020) using the special utilities provided by stdpopsim.

Estimates of average **recombination rates** for each of the chromosomes (excluding the mitochondrial genome) were taken from a recombination map inferred by Pombi et al. (2006) which itself included information from Zheng et al. (1996) (**Figure 2A**). As direct estimates of **mutation rate** (e.g., via mutation accumulation) do not currently exist for *Anopheles gambiae*, we used the genome-wide average mutation rate of $\mu = 3.5 \times 10^{-9}$ mutations per generation per site estimated by Keightley et al. (2009) for the fellow Dipteran *Drosophila melanogaster*, a rate that was used for analysis of *A. gambiae* data in Miles et al. (2017). To obtain an estimate for the default **effective population size** (N_e), we used the formula $\theta = 4\mu N_e$, with the above mutation rate ($\mu = 3.5 \times 10^{-9}$ mutations per base per generation) and a mean nucleotide diversity of $\theta \approx 0.015$, as reported by Miles et al. (2017) for the Gabon population. This resulted in an estimate of $N_e = 1.07 \times 10^6$, which we rounded down to one million. These steps were documented in the code for the stdpopsim species model, to facilitate validation and future updates. We acknowledge that some of these steps involve somewhat arbitrary choices, such as the choice of the Gabon population and rounding down of the final value. However, this should not be seen as a considerable source of misspecification, since this value of N_e is meant to provide only a rough approximation to historical population sizes and would be overwritten by a more detailed demographic model. Miles et al. (2017) inferred demographic models from *Anopheles* samples from nine different populations (locations) using the stairway plot method (Liu and Fu, 2015). We chose to include in stdpopsim the **demographic model** inferred from the Gabon sample, which consists of a single population whose size fluctuated from below 80,000 (an ancient

bottleneck roughly 10,000 generations ago) to the present-day estimate of over 4 million individuals (**Figure 2B**). To convert the timescale from generations to years, we used an **average generation time** of 1/11 years, as in Miles et al. (2017).

All of these parameters were set in the species entry in the stdpopsim catalog, accompanied by the relevant citation information, and the model underwent the standard quality-control process. The species entry may be refined in the future by adding more demographic models, updating or refining the recombination map, or updating the mutation rate estimates based on ones directly estimated for this species. Note that even if the mutation rate is updated sometime in the future, the demographic model mentioned above should still be associated with the current mutation rate ($\mu = 3.5 \times 10^{-9}$ mutations per base per generation), since this was the rate used in its inference.

Bos taurus (cattle)

Bos taurus (cattle) was added to the stdpopsim catalog during the 2020 hackathon because of its agricultural importance. Agricultural species experience strong selection due to domestication and selective breeding, leading to a reduction in effective population size. These processes, as well as admixture and introgression, produce patterns of genetic variation that can be very different from typical model species (Larson and Burger, 2013). These processes have occurred over a relatively short period of time, since the advent of agriculture roughly 10,000 years ago, and they have intensified over the years to improve food production (Gaut et al., 2018; MacLeod et al., 2013). High-quality genome assemblies are now available for several breeds of cattle (e.g., Rosen et al., 2020; Heaton et al., 2021; Talenti et al., 2022) and the use of genomic data has become ubiquitous in selective breeding (Meuwissen et al., 2001; MacLeod et al., 2014; Obšteter et al., 2021; Cesarani et al., 2022). Modern cattle have extremely low and declining genetic diversity, with estimates of effective population size around 90 in the early 1980s (MacLeod et al., 2013; VanRaden, 2020; Makanjuola et al., 2020). On the other hand, the ancestral effective population size is estimated to be roughly N_e=62,000 (MacLeod et al., 2013). This change in effective population size presents a challenge for demographic inference, selection scans, genome-wide association, and genomic prediction (MacLeod et al., 2013, 2014; Hartfield et al., 2022). For these reasons, it was useful to develop a detailed simulation model for cattle to be added to the stdpopsim catalog.

We used the most recent **genome assembly**, ARS-UCD1.2 (Rosen et al., 2020), a constant **mutation rate** of $\mu = 1.2 \times 10^{-8}$ mutations per base per generation for all chromosomes (Harland et al., 2017), and a constant **recombination rate** of $r = 9.26 \times 10^{-9}$ recombinations per base per generation for all chromosomes other than the mitochondrial genome (Ma et al., 2015). With respect to the **effective population size**, it is clear that simulating with either the ancestral or current effective population size would not generate realistic genome structure and diversity (MacLeod et al., 2013; Rosen et al., 2020). Since stdpopsim does not allow for a missing value of N_e, we chose to set the species default N_e to the ancestral estimate of 6.2×10^4. However, we strongly caution that simulating the cattle genome with any fixed value for N_e will generate unrealistic patterns of genetic variation, and recommend using a reasonably detailed demographic model. Note that the default N_e is only used in simulation if a demographic model is not specified. To this end, we implemented the **demographic model** of the Holstein breed, which was inferred by MacLeod et al. (2013) from runs of homozygosity in the whole-genome sequence of two iconic bulls. This demographic model specifies changes in the ancestral effective population size from N_e=62,000 at around 33,000 generations ago to N_e=90 in the 1980s in a series of 13 instantaneous population size changes (taken from Supplementary Table S1 in MacLeod et al., 2013). To convert the timescale from generations to years, we used an **average generation time** of 5 years (MacLeod et al., 2013). Note that this demographic model does not capture the intense selective breeding since the 1980s that has even further reduced the effective population size of cattle (MacLeod et al., 2013; VanRaden, 2020; Makanjuola et al., 2020). These effects can be modeled with downstream breeding simulations (e.g., Gaynor et al., 2020).

When setting up the parameters of the demographic model, we noticed that the inference by MacLeod et al. (2013) assumed a genome-wide fixed recombination rate of $r = 10^{-8}$ recombinations per base per generation, and a fixed mutation rate of $\mu = 9.4 \times 10^{-9}$ mutations per base per generation (considering also sequence errors). The more recently updated mutation rate assumed in the species model (1.2×10^{-8} mutations per base per generation, from Harland et al., 2017) is thus 28% higher than the rate used for inference. As a result, if genomes were simulated under this demographic model with the species' default mutation rate they would have considerably higher sequence diversity than actually observed in real genomic data. To address this, we specified a mutation rate of $\mu = 9.4 \times 10^{-9}$ in the demographic model, which then overrides the species' mutation rate when this demographic model is applied in simulation. The issue of fitting the rates used in simulation with those assumed during inference was discussed during the independent review of this demographic model, and it raised an important question about recombination rates. Since MacLeod et al. (2013) use runs of homozygosity to infer the demographic model, their results depends on the assumed recombination rate. The recombination rate assumed in inference ($r = 10^{-8}$ recombinations per base per generation) is 8% higher than the one used in the species model ($r = 9.26 \times 10^{-9}$). In its current version, stdpopsim does not allow specification of a separate recombination rate for each demographic model, so we had no simple way to adjust for this. Future versions of stdpopsim will enable such flexibility. Thus, we note that genomes simulated under this demographic model as currently implemented in stdpopsim might have slightly higher linkage disequilibrium than observed in real cattle genomes. However, we anticipate that this would affect patterns less than selection due to domestication and selective breeding, which are not yet modeled at all in stdpopsim simulations.

Conclusion

As our ability to sequence genomes continues to advance, the need for population genomic simulations of new model and non-model organisms is becoming acute. So, too, is the concomitant need for an expandable framework for implementing such simulations and guidance for how to do so. Generating realistic wholegenome simulations presents significant challenges both in coding and in choosing parameter values on which to base the simulation. With stdpopsim, we provide a resource that is uniquely poised to address these challenges as it provides easy access to state-of-the-art simulation engines and practices, and an easy procedure for including new species. Moreover, we aim for the choices regarding inclusion of new species to be driven by the needs of the population genomics community. In this manuscript we describe the expansion of stdpopsim in two ways: the addition of new features to the simulation framework that incorporate new evolutionary processes, such as non-crossover recombination, broadening the diversity of species that can be realistically modeled; and the considerable expansion of the catalog itself to include more species and demographic models.

We also formulated a series of guidelines for implementing population genomic simulations, based on insights from the community-driven process of expanding the stdpopsim catalog. These guidelines specify the basic requirements for generating a useful chromosome-level simulation for a given species, as well as the rationale behind these requirements. We also discuss special considerations for collecting relevant information from the literature, and what to do if some of that information is not available. Because this process is quite error-prone, we encourage wider adoption of "code review": researchers implementing simulations should have their parameter choices and implementation reviewed by at least one other researcher. The guidelines in this paper can be followed when implementing a simulation independently for a single study, or (as we encourage others to do) when adding code to stdpopsim, which helps to ensure its robustness and to make it available for future research. Currently, large-scale efforts such as the Earth Biogenome and its affiliated project networks are generating tens of thousands of genome assemblies. Each of these assemblies would become a candidate for inclusion into the stdpopsim catalog, although

substantial changes to the structure of stdpopsim would be required to include so many distinct species. As annotations of those genome assemblies improve over time, this information, too, can easily be added to the stdpopsim catalog.

One of the important objectives of the PopSim consortium is to leverage stdpopsim as a means to promote education and inclusion of new communities into computational biology and software development. We are keen to use outreach, such as the workshops and hackathons described here, as a way to grow the stdpopsim catalog and library while also democratizing the development of population genomic simulations in general. We predict that the increased use of chromosome-scale simulations in non-model species will lead to an improvement in inference methods, which traditionally have been quite narrowly focused on well-studied model organisms. Thus, we hope that further expansion of stdpopsim will improve the ease and reproducibility of research across a larger number of systems, while simultaneously expanding the community of software developers among population and evolutionary geneticists.

Data availability

The code for stdpopsim and the species catalog are available from: *https://github.com/popsim-consortium/stdpopsim*. The documentation page for the stdpopsim catalog is available from: *https://popsim-consortium.github.io/stdpopsim-docs/stable/catalog.html*

Acknowledgements

We wish to thank the dozens of workshop attendees, and especially the two dozen or so hackathon participants, whose combined feedback motivated many of the updates made to stdpopsim in the past two years.

Funding

M. Elise Lauterbur was supported by an NSF Postdoctoral Research Fellowship #2010884. Jean Cury was founded by DIM One Health 2017 (number RPH17094JJP) and Human Frontier Science Project, (number RGY0075/2019). David Peede is a trainee supported under the Brown University Predoctoral Training Program in Biological Data Science (NIH T32 GM128596). Per Unneberg is financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. Franz Baumdicker was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2064/1 – Project number 390727645, and EXC 2124 - Project number 390838134. Reed A. Cartwright was supported by NSF award DBI-1929850. Gregor Gorjanc was supported by the University of Edinburgh and BBSRC grant to The Roslin Institute (BBS/E/D/30002275). Ryan N. Gutenkunst was supported by NIH award R01GM127348. Jerome Kelleher was supported by the Robertson Foundation. Andrew D. Kern and Peter L. Ralph were supported by NIH award R01HG010774. Daniel R. Schrider was supported by NIH award R35GM138286

References

Adrion Jeffrey R *et al.* (2020) **A community-maintained standard library of population genetic models** *eLife* **9** https://doi.org/10.7554/eLife.54967

Amarasinghe Shanika L., Su Shian, Dong Xueyi, Zappia Luke, Ritchie Matthew E., Gouil Quentin (2020) **Opportunities and challenges in long-read sequencing data analysis** *Genome Biology* https://doi.org/10.1186/s13059-020-1935-5

Amarasinghe Shanika L, Ritchie Matthew E, Gouil Quentin (2021) **long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data** *GigaScience* **10** https://doi.org/10.1093/gigascience/giab003

Baumdicker Franz *et al.* (2021) **Efficient ancestry and mutation simulation with msprime 1.0** *Genetics* **220** https://doi.org/10.1093/genetics/iyab229

Beichman A. C., Phung T. N., Lohmueller K. E. (2017) **Comparison of Single Genome and Allele Frequency Data Reveals Discordant Demographic Histories** *G3 (Bethesda)* **7**:3605–3620

Beichman Annabel C., Huerta-Sanchez Emilia, Lohmueller Kirk E. (2018) **Using genomic data to infer historic population dynamics of nonmodel organisms** *Annu. Rev. Ecol. Evol. Syst* **49**:433–456 https://doi.org/10.1146/annurev-ecolsys-110617-062431

Benzer Seymour (1961) **On the topography of the genetic fine structure** *Proceedings of the National Academy of Sciences* **47**:403–415 https://doi.org/10.1073/pnas.47.3.403

Blischak Paul D., Barker Michael S., Gutenkunst Ryan N., Falush Daniel (2020) **Inferring the demographic history of inbred species from genome-wide SNP frequency data** *Mol. Biol. Evol* **37**:2124–2136 https://doi.org/10.1093/molbev/msaa042

Browning Sharon R., Browning Brian L. (2015) **Accurate non-parametric estimation of recent effective population size from segments of identity by descent** *The American Journal of Human Genetics* **97**:404–418 https://doi.org/10.1016/j.ajhg.2015.07.012

Cesarani A, Lourenco D, Tsuruta S, Legarra A, Nicolazzi E L, VanRaden P M, Misztal I (2022) **Multibreed genomic evaluation for production traits of dairy cattle in the United States using single-step genomic best linear unbiased predictor** *Journal of Dairy Science* **105**:5141–5152 https://doi.org/10.3168/jds.2021-21505

Chakraborty Mahul, Baldwin-Brown James G, Long Anthony D, Emerson JJ (2016) **Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage** *Nucleic acids research* **44**:e147–e147

Charlesworth Brian (2012) **The effects of deleterious mutations on evolution at linked sites** *Genetics* **190**:5–22

Cheng Shifeng *et al.* (2018) **10KP: A phylodiverse genome sequencing plan** *Gigascience* **3** https://doi.org/10.1093/gigascience/giy013

Clarkson Chris S *et al.* (2020) **Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii** *Genome research* **30**:1533–1546

Comeron Josep M (2014) **Background selection as baseline for nucleotide variation across the Drosophila genome** *PLoS Genetics* **10**

Comeron Josep M., Ratnappan Ramesh, Bailin Samuel (2012) **The many landscapes of recombination in Drosophila melanogaster** *PLoS Genet* **8**

Csilléry Katalin, Blum Michael G B, Gaggiotti Oscar E, Olivier François (2010) **Approximate Bayesian Computation (ABC) in practice** *Trends Ecol. Evol* **25**:410–8 https://doi.org/10.1016/j.tree.2010.04.001

Cury J., Haller B. C., Achaz G., Jay F. (2022) **Simulation of bacterial populations with SLiM** *Peer Community Journal* **2** https://doi.org/10.24072/pcjournal.72

Cutter A. D., Payseur B. A. (2013) **Genomic signatures of selection at linked sites: unifying the disparity among species** *Nature Reviews Genetcs* **14**:262–274 https://doi.org/10.1038/nrg3425

Darwin Tree of Life Project Consortium (2022) **Sequence locally, think globally: The Darwin Tree of Life Project** *Proceedings of the National Academy of Sciences* **119**

Dawson Elisabeth *et al.* (2002) **A first-generation linkage disequilibrium map of human chromosome 22** *Nature* **418**:544–548

Didelot X., Maiden M. C. (2010) **Impact of recombination on bacterial evolution** *Trends Microbiol* **18**:315–322

Didelot X., Meric G., Falush D., Darling A. E. (2012) **Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli** *BMC Genomics* **13**

Eldon B., Birkner M., Blath J., Freund F. (2015) **Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents?** *Genetics* **199**:841–856

Ellegren Hans (2014) **Genome sequencing and population genomics in non-model organisms** *Trends Ecol. Evol* **29**:51–63 https://doi.org/10.1016/j.tree.2013.09.008

Ellegren Hans, Smith Nick GC, Webster Matthew T (2003) **Mutation rate variation in the mammalian genome** *Current Opinion in Genetics & Development* **13**:562–568 https://doi.org/10.1016/j.gde.2003.10.008

Excoffier Laurent, Dupanloup Isabelle, Huerta-Sánchez Emilia, Sousa Vitor C., Foll Matthieu (2013) **Robust demographic inference from genomic and SNP data** *PLOS Genetics* **9**:1–17 https://doi.org/10.1371/journal.pgen.1003905

Eyre-Walker Adam, Keightley Peter D (2007) **The distribution of fitness effects of new mutations** *Nat. Rev. Genet* **8**:61061–8 https://doi.org/10.1038/nrg2146

Gaut B S, Seymour D K, Liu Q, Zhou Y (2018) **Demography and its effects on genomic variation in crop domestication** https://doi.org/10.1038/s41477-018-0210-1

Gay J., Myers S., McVean G. (2007) **Estimating meiotic gene conversion rates from population genetic data** *Genetics* **177**:881–894

Gaynor R Chris, Gorjanc Gregor, Hickey John M (2020) **AlphaSimR: an R package for breeding program simulations** *G3 Genes—Genomes—Genetics* **11** https://doi.org/10.1093/g3journal/jkaa017

Gophna U., Altman-Price N. (2022) **Horizontal Gene Transfer in Archaea-From Mechanisms to Genome Evolution** *Annu Rev Microbiol* **76**:481–502

Gower G., Picazo P. I., Fumagalli M., Racimo F. (2021) **Detecting adaptive introgression in human evolution using convolutional neural networks** *Elife* **10**

Gutenkunst Ryan N., Hernandez Ryan D., Williamson Scott H., Bustamante Carlos D. (2009) **Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data** *PLOS Genetics* **5**:1–11 https://doi.org/10.1371/journal.pgen.1000695

Haller Benjamin C., Messer Philipp W. (2019) **SLiM 3: Forward genetic simulations beyond the Wright-Fisher model** *Molecular Biology and Evolution* **36**:632–637

Harland Chad, Charlier Carole, Karim Latifa, Cambisano Nadine, Deckers Manon, Mni Myriam, Mullaart Erik, Coppieters Wouter, Georges Michel (2017) **Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle** https://doi.org/10.1101/079863

Harris K., Nielsen R. (2013) **Inferring demographic history from a spectrum of shared haplotype lengths** *PLoS Genet* **9**

Harris Kelley, Nielsen Rasmus (2016) **The genetic cost of Neanderthal introgression** *Genetics* **203**:881–891 https://doi.org/10.1534/genetics.116.186890

Hartfield M, Poulsen N Aagaard, Guldbrandtsen B, Bataillon T (2022) **Using singleton densities to detect recent selection in Bos taurus** https://doi.org/10.1002/evl3.263

Heaton Michael P *et al.* (2021) **A reference genome assembly of Simmental cattle, Bos taurus taurus** *Journal of Heredity* **112**:184–191 https://doi.org/10.1093/jhered/esab002

Hinrichs A. S. *et al.* (2006) **The UCSC Genome Browser Database: update 2006** *Nucleic Acids Res* **34**:D590–598

Howe Kevin L *et al.* (2020) **Ensembl 2021** *Nucleic Acids Research* **49**:D884–D891 https://doi.org/10.1093/nar/gkaa942

Hsieh PingHsun, Veeramah Krishna R, Lachance Joseph, Tishkoff Sarah A, Wall Jeffrey D, Hammer Michael F, Gutenkunst Ryan N (2016) **Whole genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection** *Genome Res* **26**:279–290

Hsieh PingHsun *et al.* (2021) **Evidence for opposing selective forces operating on human-specific duplicated tcaf genes in neanderthals and humans** *Nature Communications* **12**

Huber Christian D., Durvasula Arun, Hancock Angela M., Lohmueller Kirk E. (2018) **Gene expression drives the evolution of dominance** *Nat. Commun* **9** https://doi.org/10.1038/s41467-018-05281-7

Jukes T. H., Cantor C. R., Munro H.N. (1969) **Evolution of protein molecules** *Mammalian Protein Metabolism* :21–132

Keightley P. D., Trivedi U., Thomson M., Oliver F., Kumar S., Blaxter M. L. (2009) **Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines** *Genome Res* **19**:1195–1201

Kelleher Jerome, Etheridge Alison M, McVean Gilean (2016) **Efficient coalescent simulation and genealogical analysis for large sample sizes** *PLoS computational biology* **12**

Korunes Katharine L., Noor Mohamed A. F. (2017) **Gene conversion and linkage: effects on genome evolution and speciation** *Molecular Ecology* **26**:351–364 https://doi.org/10.1111/mec.13736

Kumar S., Suleski M., Craig J. M., Kasprowicz A. E., Sanderford M., Li M., Stecher G., Hedges S. B. (2022) **TimeTree 5: An Expanded Resource for Species Divergence Times** *Mol Biol Evol*

Kyriazis Christopher C., Robinson Jacqueline A., Lohmueller Kirk E. (2022) **Using computational simulations to quantify genetic load and predict extinction risk** https://doi.org/10.1101/2022.08.12.503792

Larson Greger, Burger Joachim (2013) **A population genetics view of animal domestication** *Trends in Genetics* **29**:197–205

Lewin Harris A. *et al.* (2022) **The Earth BioGenome Project 2020: Starting the clock** *Proceedings of the National Academy of Sciences* **119** https://doi.org/10.1073/pnas.2115635118

Li H., Durbin R. (2011) **Inference of human population history from individual whole-genome sequences** *Nature* **475**:493–496

Liu X., Fu Y. X. (2015) **Corrigendum: Exploring population size changes using SNP frequency spectra** *Nat Genet* **47**

Ma Li *et al.* (2015) **Cattle sex-specific recombination and genetic control from a large pedigree analysis** *PLOS Genetics* **11**:1–24 https://doi.org/10.1371/journal.pgen.1005387

Ma Xin *et al.* (2013) **Population genomic analysis reveals a rich speciation and demographic history of orang-utans (Pongo pygmaeus and Pongo abelii)** *PLoS One* **8** https://doi.org/10.1371/journal.pone.0077175

MacLeod I M, Larkin D M, Lewin H A, Hayes B J, Goddard M E (2013) **Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors** *Molecular Biology and Evolution* **30**:2209–2223 https://doi.org/10.1093/molbev/mst125

MacLeod I M, Hayes B J, Goddard M E (2014) **The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data** *Genetics* **198**:1671–1684 https://doi.org/10.1534/genetics.114.168344

Makanjuola B O, Miglior F, Abdalla E A, Maltecca C, Schenkel F S, Baes C F (2020) **Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations** https://doi.org/10.3168/jds.2019-18013

McVicker G., Gordon D., Davis C., Green P. (2009) **Widespread genomic signatures of natural selection in hominid evolution** *PLoS Genet* **5**

Meuwissen T H E, Hayes B J, Goddard M E (2001) **Prediction of total genetic value using genome-wide dense marker maps** *Genetics* **157**:1819–1829 https://doi.org/10.1093/genetics/157.4.1819

Miles A. *et al.* (2017) **Genetic diversity of the African malaria vector Anopheles gambiae** *Nature* **552**:96–100

Montano Valeria (2016) **Coalescent inferences in conservation genetics: Should the exception become the rule?** *Biol. Lett* **12** https://doi.org/10.1098/rsbl.2016.0211

Nachman Michael W. (2002) **Variation in recombination rate across the genome: Evidence and implications** *Curr. Opin. Genet. Dev* **12**:657–663 https://doi.org/10.1016/S0959-437X(02)00358-1

Navascués Miguel, Emerson Brent C (2009) **Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods** *Molecular Ecology* **18**:4390–4397

Nelson Dominic, Kelleher Jerome, Ragsdale Aaron P., Moreau Claudia, McVean Gil, Gravel Simon (2020) **Accounting for long-range correlations in genome-wide simulations of large cohorts** *PLOS Genetics* **16**:1–12 https://doi.org/10.1371/journal.pgen.1008619

Obšteter J, Jenko J, Gorjanc G (2021) **Genomic selection for any dairy breeding program via optimized investment in phenotyping and genotyping** *Frontiers in Genetics* **12** https://doi.org/10.3389/fgene.2021

Pombi March, Stump Aram D., Torre Allesandra Della, Besansky Nora J. (2006) **Variation in recombination rate across the X chromosome of Anopheles gambiae** *The American Journal of Tropical Medicine and Hygiene* **75**:901–903 https://doi.org/10.4269/ajtmh.2006.75.901

Ragsdale Aaron P., Nelson Dominic, Gravel Simon, Kelleher Jerome (2020) **Lessons learned from bugs in models of human history** *The American Journal of Human Genetics* **107**:583–588 https://doi.org/10.1016/j.ajhg.2020.08.017

Rhie Arang *et al.* (2021) **Towards complete and error-free genome assemblies of all vertebrate species** *Nature* **592**:737–746 https://doi.org/10.1038/s41586-021-03451-0

Robinson J., Kyriazis C. C., Yuan S. C., Lohmueller K. E. (2023) **Deleterious Variation in Natural Populations and Implications for Conservation Genetics** *Annu Rev Anim Biosci* **11**:93–114

Rosen Benjamin D *et al.* (2020) **De novo assembly of the cattle reference genome with single-molecule sequencing** *GigaScience* **9** https://doi.org/10.1093/gigascience/giaa021

Schiffels Stephan, Wang Ke (2020) **MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent** :147–166 https://doi.org/10.1007/978-1-0716-0199-0_7

Schrider Daniel R (2020) **Background selection does not mimic the patterns of genetic diversity produced by selective sweeps** *Genetics* **216**:499–519

Schrider Daniel R., Kern Andrew D. (2018) **Supervised machine learning for population genetics: A new paradigm** *Trends Genet* **34**:301–312 https://doi.org/10.1016/j.tig.2017.12.005

Sharakhova M. V., Hammond M. P., Lobo N. F., Krzywinski J., Unger M. F., Hillenmeyer M. E., Bruggner R. V., Birney E., Collins F. H. (2007) **Update of the Anopheles gambiae PEST genome assembly** *Genome Biol* **8**

Supek Fran, Lehner Ben (2019) **Scales and mechanisms of somatic mutation rate variation across the human genome** *DNA Repair* **81** https://doi.org/10.1016/j.dnarep.2019.102647

Talenti A *et al.* (2022) **A cattle graph genome incorporating global breed diversity** *Nature Communications* https://doi.org/10.1038/s41467-022-28605-0

Teixeira João C., Huber Christian D. (2021) **The inflated significance of neutral genetic diversity in conservation genetics** *Proc. Natl. Acad. Sci. U. S. A* **118**:1–10 https://doi.org/10.1073/pnas.2015096118

Tennessen J. A. *et al.* (2012) **Evolution and functional impact of rare coding variation from deep sequencing of human exomes** *Science* **337**:64–69

Teshima Kosuke M., Coop Graham, Przeworski Molly (2006) **How reliable are empirical genomic scans for selective sweeps?** *Genome Res* **16**:702–712 https://doi.org/10.1101/gr.5105206

Thomas C. M., Nielsen K. M. (2005) **Mechanisms of, and barriers to, horizontal gene transfer between bacteria** *Nat Rev Microbiol* **3**:711–721

VanRaden P M (2020) **Symposium review: How to implement genomic selection** *Journal of Dairy Science* **103**:5291–5301 https://doi.org/10.3168/jds.2019-17684

Wielgoss S., Barrick J. E., Tenaillon O., Cruveiller S., Chane-Woon-Ming B., Medigue C., Lenski R. E., Schneider D. (2011) **Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with Escherichia coli** *G3 (Bethesda)* **1**:183–186

Wijnker E. *et al.* (2013) **The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana** *Elife* **2**

Zheng Liangbiao, Benedict Mark Q, Cornel Anton J, Collins Frank H, Kafatos Fotis C (1996) **An integrated genetic map of the African human malaria vector mosquito, Anopheles gambiae** *Genetics* **143**:941–952

Zhou Ying, Tian Xiaowen, Browning Brian L, Browning Sharon R (2018) **POPdemog: visualizing population demographic history from simulation scripts** *Bioinformatics* **34**:2854–2855 https://doi.org/10.1093/bioinformatics/bty184

Article and author information

M. Elise Lauterbur
Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85719, USA
For correspondence: lauterbur@gmail.com
ORCID iD: 0000-0002-7362-3618

Maria Izabel A. Cavassim
Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles CA, USA
ORCID iD: 0000-0001-9726-1431

Ariella L. Gladstein
Embark Veterinary, Inc., Boston MA 02111, USA
ORCID iD: 0000-0001-7735-2336

Graham Gower
Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Denmark
ORCID iD: 0000-0002-6197-3872

Nathaniel S. Pope
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA
ORCID iD: 0000-0001-8409-7812

Georgia Tsambos
School of Mathematics and Statistics, University of Melbourne, Australia
ORCID iD: 0000-0001-7001-2275

Jeff Adrion
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA, AncestryDNA, San Francisco CA 94107, USA
ORCID iD: 0000-0003-1021-6000

Saurabh Belsare
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA
ORCID iD: 0000-0002-8148-1867

Arjun Biddanda
54Gene, Inc., Washington DC 20005, USA
ORCID iD: 0000-0003-1861-1523

Victoria Caudill
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA
ORCID iD: 0000-0002-0577-5513

Jean Cury
Université Paris-Saclay, CNRS, INRIA, Laboratoire Interdisciplinaire des Sciences du Numérique, UMR 9015 Orsay, France
ORCID iD: 0000-0002-6462-8783

Ignacio Echevarria
School of Life Sciences, University of Glasgow, Glasgow, UK
ORCID iD: 0009-0000-5158-709X

Benjamin C. Haller
Department of Computational Biology, Cornell University, Ithaca NY, USA
ORCID iD: 0000-0003-1874-8327

Ahmed R. Hasan
Department of Cell and Systems Biology, University of Toronto, Toronto ON, Canada, Department of Biology, University of Toronto Mississauga, Mississauga ON, Canada
ORCID iD: 0000-0003-0002-8399

Xin Huang
Department of Evolutionary Anthropology, University of Vienna, Vienna, Austria, Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Vienna, Austria
ORCID iD: 0000-0002-9918-9602

Leonardo Nicola Martin Iasi
Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
ORCID iD: 0009-0005-8020-0380

Ekaterina Noskova
Computer Technologies Laboratory, ITMO University, St Petersburg, Russia
ORCID iD: 0000-0003-1168-0497

Jana Obšteter
Agricultural Institute of Slovenia, Department of Animal Science, Ljubljana, Slovenia
ORCID iD: 0000-0003-1511-3916

Vitor Antonio Corrêa Pavinato
Entomology Department, The Ohio State University, Wooster OH, USA
ORCID iD: 0000-0003-2483-1207

Alice Pearson
Department of Genetics, University of Cambridge, Cambridge, UK, Department of Zoology, University of Cambridge, Cambridge, UK

David Peede
Department of Ecology, Evolution, and Organismal Biology, Brown University, Providence RI, USA, Center for Computational Molecular Biology, Brown University, Providence RI, USA
ORCID iD: 0000-0002-4826-0464

Manolo F. Perez
Department of Genetics and Evolution, Federal University of Sao Carlos, Sao Carlos 13565905, Brazil
ORCID iD: 0000-0002-4642-7793

Murillo F. Rodrigues
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA
ORCID iD: 0000-0001-7508-1384

Chris C. R. Smith
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA
ORCID iD: 0000-0002-6470-3413

Jeffrey P. Spence
Department of Genetics, Stanford University School of Medicine, Stanford CA 94305, USA
ORCID iD: 0000-0002-3199-1447

Anastasia Teterina
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA
ORCID iD: 0000-0003-3016-8720

Silas Tittes
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA
ORCID iD: 0000-0003-4697-7434

Per Unneberg
Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Husargatan 3, SE-752 37 Uppsala, Sweden
ORCID iD: 0000-0001-5735-3315

Juan Manuel Vazquez
Department of Integrative Biology, University of California, Berkeley, Berkeley CA, USA
ORCID iD: 0000-0001-8341-2390

Ryan K. Waples
Department of Biostatistics, University of Washington, Seattle WA, USA
ORCID iD: 0000-0003-0526-6425

Anthony Wilder Wohns
Broad Institute of MIT and Harvard, Cambridge MA 02142, USA
ORCID iD: 0000-0001-7353-1177

Yan Wong
Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, UK
ORCID iD: 0000-0002-3536-6411

Franz Baumdicker
Cluster of Excellence - Controlling Microbes to Fight Infections, Eberhard Karls Universität Tübingen, Tübingen, Baden-Württemberg, Germany
ORCID iD: 0000-0001-9106-7259

Reed A. Cartwright
School of Life Sciences and The Biodesign Institute, Arizona State University, Tempe AZ, USA
ORCID iD: 0000-0002-0837-9380

Gregor Gorjanc
The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh EH25 9RG, UK
ORCID iD: 0000-0001-8008-2787

Ryan N. Gutenkunst
Department of Molecular and Cellular Biology, University of Arizona, Tucson AZ 85721, USA
ORCID iD: 0000-0002-8659-0579

Jerome Kelleher
Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, UK
ORCID iD: 0000-0002-7894-5253

Andrew D. Kern
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA
ORCID iD: 0000-0003-4381-4680

Aaron P. Ragsdale
Department of Integrative Biology, University of Wisconsin-Madison, Madison WI, USA
ORCID iD: 0000-0003-0715-3432

Peter L. Ralph
Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA, Department of Mathematics, University of Oregon, Eugene OR 97402, USA
ORCID iD: 0000-0002-9459-6866

Daniel R. Schrider
Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill NC 27599, USA
ORCID iD: 0000-0001-5249-4151

Ilan Gronau
Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel
For correspondence: ilan.gronau@runi.ac.il
ORCID iD: 0000-0001-8536-4062

Copyright

Editors

Reviewing Editor
Ziyue Gao
University of Pennsylvania, Philadelphia, United States of America

Senior Editor
Molly Przeworski
Columbia University, New York, United States of America

Reviewer #1 (Public Review):

stdpopsim is an existing, community-driven resource to support population genetics simulations across multiple species. This paper describes improvements and extensions to this resource and discusses various considerations of relevance to chromosome-scale evolutionary simulations. As such, the paper does not analyse data or present new results but

rather serves as a general and useful guide for anyone interested in using the stdpopsim resource or in population genetics simulations in general.

Two new features in stdpopsim are described, which expand the types of evolutionary processes that can be simulated. First, the authors describe the addition of the ability to simulate non-crossover recombination events, i.e. gene conversion, in addition to standard crossover recombination. This will allow for simulations that come closer to the actual recombination processes occurring in many species. Second, the authors mention how genome annotations can now be incorporated into the simulations, to allow different processes to apply to different parts of the genome - however, the authors note that this addition will be further detailed in a separate, future publication. These additions to stdpopsim will certainly be useful to many users and represent a step forward in the degree of ambition for realistic population genetics simulations.

The paper also describes the expansion of the community-curated catalog of pre-defined, ready-to-use simulation set-ups for various species, from the previous 6 to 21 species (though not all new species have demographic models implemented, some have just population genetic parameters such as mutation rates and generation times). For each species, an attempt was made to implement parameters and simulations that are as realistic as possible with respect to what's known about the evolutionary history of that species, using only information that can be traced to the published literature. This process by which this was done appears quite rigorous and includes a quality-control process involving two people. Two examples are given, for Anopheles gambiae and Bos taurus. The detailed discussion of how various population genetic and demographic parameters were extracted from the literature for these two species usefully highlights the numerous non-trivial steps involved and showcases the great deal of care that underlies the stdpopsim resource.

The paper is clearly written and well-referenced, and I have no technical or conceptual concerns. The paper will be useful to anyone interested in population genetics simulations, and will hopefully serve as an inspiration for the broader effort of making simulations increasingly more realistic and flexible, while at the same time trying to make them accessible not just to a small number of experts.

- https://doi.org/10.7554/eLife.84874.2.sa3

Reviewer #2 (Public Review):

Lauterbur et al. present a description of recent additions to the stdpopsim simulation software for generating whole-genome sequences under population genetic models, as well as detailed general guidelines and best practices for implementing realistic simulations within stdpopsim and other simulation software. Such realistic simulations are critical for understanding patterns in genetic variation expected under diverse processes for study organisms, training simulation-intensive models (e.g., machine learning and approximate Bayesian computation) to make predictions about factors shaping observed genetic variation, and for generating null distributions for testing hypotheses about evolutionary phenomena. However, realistic population genomic simulations can be challenging for those who have never implemented such models, particularly when different evolutionary parameters are taken from a variety of literature sources. Importantly, the goal of the authors is to expand the inclusivity of the field of population genomic simulation, by empowering investigators, regardless of model or non-model study system, to ultimately be able to effectively test hypotheses, make predictions, and learn about processes from simulated genomic variation. Continued expansion of the stdpopsim software is likely to have a significant impact on the evolutionary genomics community.

Strengths:

This work details an expansion from 6 to 21 species to gain a greater breadth of simulation capacity across the tree of life. Due to the nature of some of the species added, the authors implemented finite-site substitution models allowing for more than two allelic states at loci, permitting proper simulations of organisms with fast mutation rates, small genomes, or large effect sizes. Moreover, related to some of the newly added species, the authors incorporated a mechanism for simulating non-crossover recombination, such as gene conversion and horizontal gene transfer between individuals. The authors also added the ability to annotate and model coding genomic regions.

In addition to these added software features, the authors detail guidelines and best practices for implementing realistic population genetic simulations at the genome-scale, including encouraging and discussing the importance of code review, as well as highlighting the sufficient parameters for simulation: chromosome level assembly, mean mutation rate, mean recombination rate or recombination map if available, effective size or more realistic demographic model if available, and mean generation time. Much of these best practices are commonly followed by population genetic modelers, but new researchers in the field seeking to simulate data under population genetic models may be unfamiliar with these practices, making their clear enumeration (as done in this work) highly valuable for a broad audience. Moreover, the mechanisms for dealing with issues of missing parameters discussed in this work are particularly useful, as more often than not, estimates of certain model parameters may not be readily available from the literature for a given study system.

Weaknesses:

An important update to the stdpopsim software is the capacity for researchers to annotate coding regions of the genome, permitting distributions of fitness effects and linked selection to be modeled. However, though this novel feature expands the breadth of processes that can be evaluated as well as is applicable to all species within the stdpopsim framework, the authors do not provide significant detail regarding this feature, stating that they will provide more details about it in a forthcoming publication. Compared to this feature, the additions of extra species, finite-site substitution models, and non-crossover recombination are more specialized updates to the software.

- https://doi.org/10.7554/eLife.84874.2.sa2

Reviewer #3 (Public Review):

Lauterbur et al. present an expansion of the whole-genome evolution simulation software "stdpopsim", which includes new features of the simulator itself, and 15 new species in their catalog of demographic models and genetic parameters (which previously had 6 species). The list of new species includes mostly animals (12), but also one species of plant, one of algae, and one of bacteria. While only five of the new animal species (and none of the other organisms) have a demographic model described in the catalog, those species showcase a variety of demographic models (e.g. extreme inbreeding of cattle). The authors describe in detail how to go about gathering genetic and demographic parameters from the literature, which is helpful for others aiming to add new species and demographic models to the stdpopsim catalog. This part of the paper is the most widely relevant not only for stdpopsim users but for any researcher performing population genomics simulations. This work is a concrete contribution towards increasing the number of users of population genomic simulations and improving reproducibility in research that uses this type of simulations.

- https://doi.org/10.7554/eLife.84874.2.sa1

Author Response:

The following is the authors' response to the original reviews.

We are very glad that the editor and reviewers found our paper of broad interest to the community of population, evolutionary, and ecological genetics. We thank them for their positive feedback and insightful comments and suggestions. We have revised our manuscript to address some of the issues raised by the review. The main change we made was providing a detailed discussion of limitations of simulated genomes, focusing on considerations one needs to make when selecting a demographic model. This can be found in a new section "Limitations of simulated genomes" (pages 9-10). We made a few additional adjustments in other parts of the text based on the reviewers' suggestions. They are all listed in the detailed point-by-point response to reviewers comments and questions below.

> *Editor:*
>
> 1. *It was noted that demographic models (or genomic parameters) that are inferred based on certain aspects of the genomic data (eg., site frequency spectrum, haplotype structure) may not recapitulate other aspects of the data. In other words, any inferred demographic models are expected to reliably reproduce only some aspects of the genetic variation data but not necessarily all. It would be helpful to emphasize this limitation in the manuscript and to include a table summarizing the types of variation that the demographic models for the catalogued species were based on.*

This is a very important point, which we addressed in the revision by adding a section entitled "Limitations of simulated genomes". This section discusses the considerations that one should make when selecting an inferred demographic model to implement in simulation. This includes the samples used in analysis, the method used for inference, as well as various filters. In this section we also point to the documentation page of the stdpopsim catalog, which provides information about each demographic model that can help users decide whether it is appropriate for their needs. We decided not to summarize this information in a succinct table in the manuscript because it is not straightforward to summarize the strengths and potential limitations of each model in a table. Instead, we will expand the summary provided for each demographic model in the documentation page to provide additional information. See response to the second reviewer's comment on this topic for more details.

> 1. *It will make stdpopsim more user-friendly to include an automated module that can visualize a demographic model given the corresponding parameters (or simulation scripts).*

As mentioned in the response to the first reviewer's comment on this subject, the documentation page of the stdpopsim catalog provides a brief summary for each demographic model, including a graphical representation. See response below for more details.

> ***Reviewer #1:***
>
> *In the introduction, the authors cite numerous efforts to generate high-quality reference genomes. That's not an issue in itself, but leading with this might send the message to some readers that it is these reference genome efforts that are driving the need for population genomics analysis and simulation tools, which is not really the case - why not instead give some citation attention to actual population genomics projects aiming to address the types of evolutionary questions this paper is concerned with? The reference*

> *genome citations would fit better in the section dealing with reference genomes, where they already appear.*

Indeed, the desire to answer complex evolutionary questions is the main motivation for sequencing these genomes and also for generating realistic genome simulations. The reason we chose to lead with the genome-sequencing efforts is that high quality genome data is an important prerequisite for obtaining parameters for chromosome-scale simulations. So, with that perspective, these efforts which we cite are the driving force behind expansion of stdpopsim in the near future. Thus, we decided to leave these citations in the introduction. To balance things out, we now start the introduction with a statement about board questions in population genetics. Moreover, after we list the genome sequencing efforts, we added a list of specific types of questions that can be addressed by these newly emerging genomes, with relevant citations. The beginning of the introduction now reads:

"Population genetics allows us to answer questions across scales from deep evolutionary time to ongoing ecological dynamics, and dramatic reductions in sequencing costs enable the generation of unprecedented amounts of genomic data that can be used to address these questions (Ellegren, 2014). Ongoing efforts to systematically sequence life on Earth by initiatives such as the Earth Biogenome (Lewin et al., 2022) and its affiliated project networks, such as Vertebrate Genomes (Rhie et al., 2021), 10,000 Plants (Cheng et al., 2018) and others (Darwin Tree of Life Project Consortium, 2022), are providing the backbone for enormous increases in the amount of population-level genomic data available for model and non-model species. These data are being used, among other things, in inference of population history and demographic parameters (Beichman et al., 2018), studying adaptive introgression (Gower et al., 2021), distinguishing adaptation from drift (e.g. Hsieh et al., 2021), and understanding the implications of deleterious variation in populations of conservation concern (e.g. Robinson et al., 2023)."

> *Something that would be useful for the stdpopsim resource in general, though not necessarily something for the paper, would be some kind of more human-friendly representation of the demographic models implemented in the curated library. Perhaps I'm not looking in the right place, but as far as I can tell, if I want to study the curated demographic models, I need to go into the Python scripts on the stdpopsim GitHub page (e.g.*
>
> *https://github.com/popsim-consortium/stdpopsim/tree/main/stdpopsim/catalog/BosTau). Here the various parameters and demographic events are hard-coded into the scripts. To understand the model being implemented, one thus needs to go dig into these scripts - something which is not necessarily very accessible to all researchers. Visual representations, such as the one for Anopheles gambiae in Fig 2. in the paper, are more widely accessible. I wonder if such figures could be produced for all the curated models and included in the GitHub folders alongside the scripts, perhaps aided by an existing model visualization software such as POPdemog. Again, I would not suggest that this is necessary for the paper, but if practically feasible I think it would be a useful addition to the resource in the longer term.*

This is a very good point. The stdpopsim catalog actually has a documentation page that provides a brief summary for each demographic model, including a graphical representation. This graphical representation is generated using demesdraw applied to the demographic model object implemented in the code. Thus, potential users do not have to dig through the Python code to figure out the details of the demographic model. We used a similar approach to generate the image of the demographic history of *A. gambiae* for Fig. 2 of the paper. The documentation page is an important part of the stdpopsim catalog, and we now added a link to it in section "Data availability", and we mention it in key places in the manuscript, such as the caption of Fig 2.

> ***Reviewer #2:***
>
> *An important update to the stdpopsim software is the capacity for researchers to annotate coding regions of the genome, permitting distributions of fitness effects and linked selection to be modeled. However, though this novel feature expands the breadth of processes that can be evaluated as well as is applicable to all species within the stdpopsim framework, the authors do not provide significant detail regarding this feature, stating that they will provide more details about it in a forthcoming publication. Compared to this feature, the additions of extra species, finite-site substitution models, and non-crossover recombination are more specialized updates to the software.*
>
> *It would be helpful to provide additional information regarding the coding annotation (and associated distribution of fitness effects and linked selection) that is implemented in the current version of stdpopsim, but will be detailed in a forthcoming paper. This is not to take away from the forthcoming paper, but I believe this is the most important update to the software, and the current manuscript only brushes over it.*

We agree that implementation of selection in simulations is a significant addition to stdpopsim. However, our intention in this manuscript is to focus on the separate effort we made in the last two years to expand the utility of stdpopsim to a more diverse set of species. We think the manuscript stands firmly even without discussing in detail the new features that allow modeling selection. The main reason we briefly mention these features in sections "Additions to stdpopsim" and "Basic setup for chromosome-level simulations" is because the released version of stdpopsim contains implemented DFEs for a few species, and we did not want to completely ignore this. We thus added a brief comment at the end of the "Basic setup" section (page 8) mentioning the three model species for which the stdpopsim catalog currently has annotations and implemented DFE models. We think that a more detailed description of how these features and how they should be used is best left to the manuscript that the PopSim community is currently writing (preprint expected later this year).

> *When it comes to simulating realistic genomic data, the authors clearly lay out that parameters obtained from the literature must be compatible, such as the same recombination and mutation rates used to infer a demographic history should also be used within stdpopsim if employing that demographic history for simulation. This is a highly important point, which is often overlooked. However, it is also important that readers understand that depending on the method used to estimate the demographic history, different demographic models within stdpopsim may not reproduce certain patterns of genetic variation well. The authors do touch on this a bit, providing the example that a constant size demographic history will be unable to capture variation expected from recent size changes (e.g., excess of low-frequency alleles). However, depending on the data used to estimate a demographic history, certain types of variation may be unreliably modeled (Biechman et al. 2017; G3, 7:3605-3620). For example, if a site frequency spectrum method was used to estimate a demographic history, then the simulations under this model from y stdpopsim may not recapitulate the haplotype structure well in the observed species. Similarly, if a method such as PSMC applied to a single diploid genome was used to estimate a demographic history, then the simulations under this model from stdpopsim may not recapitulate the site frequency spectrum well in the observed species. Though the authors indicate that citations are given to each demographic model and model parameter for each species, this may not be sufficient for a novice researcher in this field to understand what forms of genomic variation the models may be capable of reliably producing. A potential worry is that the inclusion of a species within stdpopsim may serve as an endorsement to users regarding the available simulation models (though I understand this is not the case by the authors), and it would be helpful if users and readers were guided on the type of variation the models should be*

> *able to reliably reproduce for each species and demographic history available for each species. It would be helpful to include a table with types of observed variation that the current set of 21 species (and associated demographic histories) are likely and unlikely to recapitulate well.*

This is a very important point, which we now address in the section "Limitations of simulated genomes", which we added to the manuscript. In this section, we expand on this topic and discuss various things that will affect the way simulated genomes reflect true sequence variation. This includes the choice of demographic inference method, but also the analyzed samples, and various filters. The main message of this section is that one should consider various things when deciding to implement a demographic model in simulation (or selecting a model among those implemented in stdpopsim). We also cite studies (including Beichman, et al. 2017), which compared different approaches to demography inference. However, we note that the conclusions of these comparisons are not as straightforward as the reviewer suggests. In particular, methods that make use of the site frequency spectrum (such as dadi) should be able to capture some aspects of haplotype structure, because this information is encoded in the demographic history. Furthermore, a demographic history inferred from a single genome (e.g., using PSMC) should do a reasonable job approximating some aspects of the site frequency spectrum. In other words, the aspects of genetic variation not modeled well by a given demographic inference method are not always predicted in a straightforward way. This is why we avoid summarizing this information in a table in the manuscript. The 2nd paragraph of the "Limitations of simulated genomes" section addresses some of these subtle considerations. In particular, we suggest that considering a demographic model for simulation requires some familiarity with the inference method and the way it was applied to data. Regarding the demographic models currently implemented in stdpopsim, we provide some information about each model in the documentation page of the catalog. When selecting a demographic model from the catalog, users should make use of this documentation to guide their decision. This is mentioned in the 3rd paragraph of the "Limitations of simulated genomes" section. Following-up on this issue, we intend to review the documentation and make sure it provides sufficient information for each demographic model. See this GitHub issue.

> ***Reviewer #3:***
>
> *- p5, 2nd paragraph: I think many Biologists, myself included, will think of horizontal gene transfer mostly as plasmids being transferred among bacteria and adding extra genetic material, not as homologous bacterial recombination. This made me confused about modelling horizontal gene transfer in the same way as gene conversion. It may be helpful for some readers if you specify that you are modelling this particular type of horizontal gene transfer. Some explanation along the lines of what is in Cury et al (2022) would be enough.*

This is a good point. We modified the text in that sentence in the 2nd paragraph on page 5 to clarify that we are modeling non-crossover homologous recombination, and not incorporation of exogenous DNA (e.g., via plasmid transfer). The relevant part of the text now says:

"In bacteria and archaea, genetic material can be exchanged through horizontal gene transfer, which can add new genetic material (e.g., via the transfer of plasmids) or replace homologous sequences through homologous recombination (Thomas and Nielsen, 2005; Didelot and Maiden, 2010; Gophna and Altman-Price, 2022). However, the initial version of stdpopsim used crossover recombination to stand in for these processes. Although we cannot currently simulate varying gene content (as would be required to simulate the addition of new genetic material by horizontal gene transfer), the msprime and SLiM simulation engines

now allow gene conversion, which has the same effect as non-crossover homologous recombination.

Following (Cury et al., 2022), we use this to include non-crossover homologous recombination in bacterial and archaeal species."

> *- p5, 3rd paragraph: When you say gene conversion is turned off by default, you could refer to table 1 and briefly mention the consequence of ignoring gene conversion.*

We agree that it is important to note that avoiding to model gene conversion may lead to faulty lengths of shared haplotypes across individuals. This is implied by the statement we make in the beginning of the 3rd paragraph on page 5, where we lay out the motivation for modeling gene conversion in simulation. Following the reviewer's suggestion, we now added a statement about this in the end of that paragraph:

"Note that ignoring gene conversion may result in a slightly skewed distribution of shared haplotypes between individuals (see Table 1)"

> *- p7, item 1 and p9, 1st paragraph: I am not sure what you mean by genetic map here, can you define this term? I am not sure if it is synonymous with gene annotations, a recombination map, or something else. The linkage map doesn't seem to make sense to me here.*

The term 'genetic map' referred to the recombination map whenever it was used in the manuscript. To avoid any confusion, we now removed all mentions of 'genetic map', and use 'recombination map' instead. The recombination map is relevant in item 1 of page 7 because in species with poor assemblies you will not be able to reliably estimate recombination maps, making chromosome-scale simulations less effective. In the 1st paragraph of page 9, we discuss the issue of lifting over coordinates from one assembly to another, and if you have a recombination map estimated in one assembly, you might need to lift it over to another assembly to apply it in your simulation.

> *- Table 1, last row, middle column: when you say "simulated population", I think it is a bit ambiguous. You mean "the true population that we are trying to simulate", but could be read as "the population data that was generated by simulation". I would delete the word simulated here.*

What we mean here is that the selected effective population size should reflect the observed genetic diversity in real genomic data. We realize that the previous wording was confusing, and changed this to the following:

"Set the effective population size (Ne) to a value that reflects the observed genetic diversity"

> *- Figure 2, and other places when you refer to mutation and recombination rate (eg p11, last paragraph), can you include the units (e.g. per base pair, per generation)?*

Throughout the manuscript, rates are always specified per base per generation. In Figure 2, this is specified in the caption (3rd line). We added units in other places in section "Examples of added species" on pages 12-13, where they were indeed missing.

> *- p11, "default effective population size": can you use a more descriptive word instead of the default? Maybe the historical average? Also, what is this value used for in the simulations when there is a demographic model specified (as in the case of Anopheles)?*

We think that “default effective population size” is the most appropriate term to use here, since we are referring to the parameter in the species model in stdpopsim. It is correct that the value of this parameter should reflect the historical average size in some sense, but it is really unclear what this should be in the case of a species like *Bos taurus*, which experienced a very dramatic bottleneck in the recent past. We address this subtle, yet important, issue in the sentence preceding this one. If a demographic model is specified in simulation, it overrides the default effective population size, and its value is ignored (which is why we refer to it as ‘default’). We added a short sentence clarifying this in the 2nd paragraph of the “*Bos Taurus*” section (now page 12).

“Note that the default Ne is only used in simulation if a demographic model is not specified.”

> *- p8, when you say "Such simulations are useful for a number of purposes, but they cannot be used to model the influence of natural selection on patterns of genetic variation.": You may want to bring up the discussion that many of these neutral parameters taken from the literature could have been estimated assuming genome-wide neutrality, and thus ignoring the effect of background selection. Therefore the parameter values might reflect some effect of background selection that was unaccounted for during their estimation.*

This is an important subtle point, which we now address in the section “Limitations of simulated genomes”, which we added to the revised manuscript. In that section, we discuss various limitations of simulations, focusing on inferred demographic models. We address the potential influence of the segments selected for analysis toward the end of 2nd paragraph in that section (page 9):

“... all methods assume that the input sequences are neutrally evolving. This implies that technical choices, such as the specific genomic segments analyzed and various filters, may also influence the inferred model and its ability to model observed genetic variation.”

Interestingly, background selection in itself typically does not have a strong effect on the inferred model. This is something that is examined in the forthcoming publication that presents simulations with natural selection in stdpopsim.

> *- Why are some concepts written in bold (eg effective population size, demographic model)? Were you planning to make a vocabulary box? I think this is a good idea given that you are aiming for a public that can include people who are not very familiar with some population genetics concepts.*

In the “Examples of added species” section, we use boldface fonts to highlight the model parameters that were determined for each species. We added a statement clarifying this in the beginning of this section (page 11), and made sure that all the relevant parameters were consistently highlighted throughout this section. In other sections, we use boldface fonts only for titles. A few cases that did not conform to this rule were removed in the current version. We did not intend on adding a vocabulary box, but considered this when revising the manuscript, due to the reviewer’s suggestion. However, we found it difficult to converge on a small (yet comprehensive) set of terms with accurate and succinct definitions. We think that the important terms are adequately defined within the text of the manuscript, providing sufficient information also for readers who are not expert population geneticists.

> *- p4, 2nd paragraph: Are these automated scripts that are used to compare models publicly available? If you are suggesting that people use this approach generally when coming up with a simulation model (p8, penultimate paragraph), it would be helpful to have access to these automated scripts.*

The scripts are part of the public stdpopsim repository on GitHub, and may be used by anyone. Some components of these scripts are more easy to apply in general, such as comparing a demographic model with one implemented separately by the reviewer. This step, for example, is achieved by application of the Demography.is_equivalent method in msprime. Other parts of the comparison depend on the specific structure of python objects used by stdpopsim, so they are not likely to be useful when implementing simulations outside the framework of stdpopsim.

> *- p9, 1st paragraph, and p.12 2nd paragraph: instead of adjusting the mutation rate to fit the demographic model (and using an old estimate of the mutation rate), would it be ok to adjust the demographic model to fit the new mutation rate? E.g. with a new mutation rate that is the double of a previous estimate, would it be ok to just divide Ne by 2 such that Ne*mu is constant (in a constant population size model)? I imagine this could get complicated with population size changes.*

In principle, this could be done if you were simulating neutrally evolving sequences (without modeling natural selection). Since the coalescence is scale-free, then you can scale down all population sizes and divergence times by a multiplicative factor, and scale up migration rates and the mutation rate by the same factor, and you get the exact same distribution over the output sequences. However, making sure you get the scaling right is tricky and is quite error-prone. Especially considering the fact that you have to do this every time the mutation rate of a species is updated. Moreover, once you start modeling natural selection, this scale-free property no longer holds. Thus, the simple solution we came up with in stdpopsim is to attach to each demographic model the mutation rate used in its inference.