eLife

Genetics and Genomics

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

**Reviewed Preprint posted**
3 March 2023

**Sent for peer review**
14 November 2022

**Posted to bioRxiv**
31 October 2022

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

**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 simple 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

Dramatic reductions in sequencing costs are enabling the generation of unprecedented amounts of genomic data for a huge variety of species (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 populationlevel genomic data available for model and non-model species. These data are being used to answer questions across scales from deep evolutionary time to ongoing ecological dynamics. Methods that use these data for purposes such as inference of demographic history and natural selection are also flourishing (Beichman et al.,2018). While past methods development focused on humans and a few key model systems such as *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.

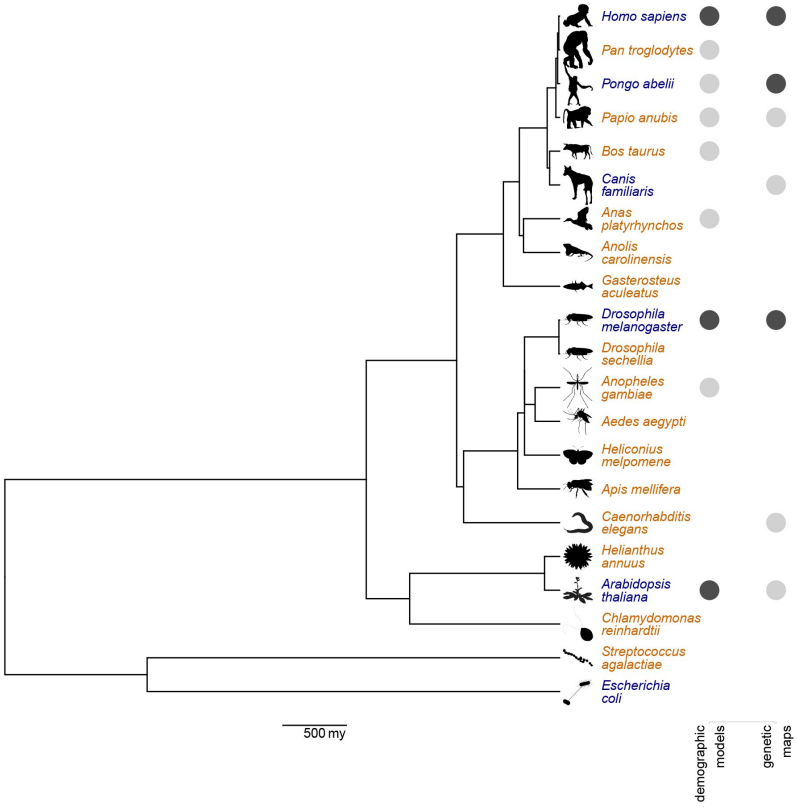


**Figure 1:**

Phylogenetic tree of species available in the stdpopsim catalog, including the six species we published in the original release (Adrion et al., 2020; Kumar et al., 2022). The horizontal bar below the tree indicates 500 million years (my).

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 genetic maps or demographic models (see Figure 1 ), but this lays a framework for future contributions.

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, recombination occurs through horizontal transfer of homologous DNA segments from one organism to another (Thomas and Nielsen, 2005; Didelot and Maiden, 2010; Gophna and Altman-Price,2022). As a result, such species cannot be realistically simulated with a recombination model that considers only crossovers, as did the

initial version of stdpopsim. To address this, we made use of features of the msprime and SLiM simulation engines for modeling non-crossover recombination (Cury et al., 2022). Modeling recombination in a bacterial or archaeal species in stdpopsim is done 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 $8.9 \times 10^{-11}$ 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 ($8.9 \times 10^{-11}$) 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.

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.

# 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 estimates of recombination rate, genetic 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. 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)).

## 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 genetic maps and genome annotations to the coordinates of the current genome assembly (Hinrichs et al., 2006).

## 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.

| 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 average observed genetic diversity in the simulated population | 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.

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

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 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.

## *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 ($\mu = 3.5 \times 10^{-9}$) and a mean nucleotide diversity of $\theta \approx 0.015$, as reported by Mileset 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 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)).

**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**

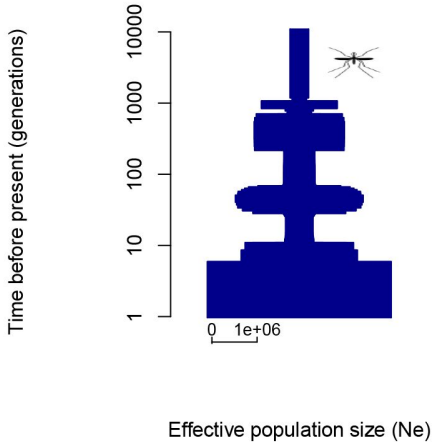


**Figure 2:**

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 $N_e = 10^6$.

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 ($\mu = 3.5 \times 10^{-9}$), 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**$\mu = 1.2 \times 10^{-8}$ for all chromosomes (Harland et al., 2017), and a constant **recombination rate**$r = 9.26 \times 10^{-9}$ 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); Rosenet 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 $6.2 \times 10^4$. 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. 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 $r = 10^{-8}$, and a fixed mutation rate $\mu = 9.4 \times 10^{-9}$ (considering also sequence errors). The more recently updated mutation rate assumed in the species model ($1.2 \times 10^{-8}$ 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

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 ($r = 10^{-8}$) 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

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

All resources are available from https://github.com/popsim-consortium/stdpopsim

## 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 , Cole Christopher B , Dukler Noah , Galloway Jared G , Gladstein Ariella L , Gower Graham , Kyriazis Christopher C , Ragsdale Aaron P , Tsambos Georgia , Baumdicker Franz , Carlson Jedidiah , Cartwright Reed A , Durvasula Arun , Gronau Ilan , Kim Bernard Y , McKenzie Patrick , Messer Philipp W , Noskova Ekaterina , Vecchyo Diego Ortega-Del , Racimo Fernando , Struck Travis J , Gravel Simon , Gutenkunst Ryan N , Lohmueller Kirk E , Ralph Peter L , Schrider Daniel R , Siepel Adam , Kelleher Jerome , Kern Andrew D (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* **21:** 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 10:** https://doi.org/10.1093/gigascience/giab003

Baumdicker Franz , Bisschop Gertjan , Goldstein Daniel , Gower Graham , Ragsdale Aaron P , Tsambos Georgia , Zhu Sha , Eldon Bjarki , Ellerman E Castedo , Galloway Jared G , Gladstein Ariella L , Gorjanc Gregor , Guo Bing , Jeffery Ben , Kretzschumar Warren W , Lohse Konrad , Matschiner Michael , Nelson Dominic , Pope Nathaniel S , Quinto-Cortés Consuelo D , Rodrigues Murillo F , Saunack Kumar , Sellinger Thibaut , Thornton Kevin , van Kemenade Hugo , Wohns Anthony W , Wong Yan , Gravel Simon , Kern Andrew D , Koskela Jere , Ralph Peter L , Kelleher Jerome (2021) **Efficient ancestry and mutation simulation with msprime 1.0 220:** https://doi.org/10.1093/genetics/iyab229

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 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 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 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 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 44:**

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

Cheng Shifeng , Melkonian Michael , Smith Stephen A. , Brockington Samuel , Archibald John M. , Delaux Pierre-Marc , Li Fay-Wei , Melkonian Barbara , Mavrodiev Evgeny V. , Sun Wenjing , Fu Yuan , Yang Huanming , Soltis Douglas E. , Graham Sean W. , Soltis Pamela S. , Liu Xin , Xu Xun , Wong Gane Ka-Shu (2018) **10kp: A phylodiverse genome sequencing plan 3:** https://doi.org/10.1093/gigascience/giy013

Clarkson Chris S , Miles Alistair , Harding Nicholas J , Lucas Eric R , Battey CJ , Amaya-Romero Jorge Edouardo , Kern Andrew D , Fontaine Michael C , Donnelly Martin J , Lawniczak Mara KN (2020) **Genome variation and population structure among 1142 mosquitoes of the african malaria vector species anopheles gambiae and anopheles coluzzii 30:**1533–1546

Comeron J. M. , Ratnappan R. , Bailin S. (2012) **The many landscapes of recombination in Drosophila melanogaster 8:**

Comeron Josep M (2014) **Background selection as baseline for nucleotide variation across the drosophila genome 10:**

Csilléry Katalin , Blum Michael G B , Gaggiotti Oscar E , François Olivier (2010) **Approximate Bayesian Computation (ABC) in practice 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 CommunityJournal* **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 14:**262–274
https://doi.org/10.1038/nrg3425

undefined undefined (2022) **Sequence locally, think globally: The Darwin Tree of Life Project 119:**

Dawson Elisabeth , Abecasis Gonçalo R , Bumpstead Suzannah , Chen Yuan , Hunt Sarah , Beare David M , Pabial Jagjit , Dibling Thomas , Tinsley Emma , Kirby Susan (2002) **A first-generation linkage disequilibrium map of human chromosome 22 418:**544–548

Didelot X. , Maiden M. C. (2010) **Impact of recombination on bacterial evolution 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:**256

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

Ellegren Hans (2014) **Genome sequencing and population genomics in non-model organisms 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 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 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 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 177:**881–894

Gaynor R Chris , Gorjanc Gregor , Hickey John M (2020) **AlphaSimR: an R package for breeding program simulations 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

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 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 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 Kelley , Nielsen Rasmus (2016) **The genetic cost of neanderthal introgression 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 , Smith Timothy P L , Bickhart Derek M , Ley Brian L Vander , Kuehn Larry A , Oppenheimer Jonas , Shafer Wade R , Schuetze Fred T , Stroud Brad , McClure Jennifer C , Barfield Jennifer P , Blackburn Harvey D , Kalbfleisch Theodore S , Davenport Kimberly M , Kuhn Kristen L , Green Richard E , Shapiro Beth , Rosen Benjamin D (2021) **A Reference Genome Assembly of Simmental Cattle, Bos taurus taurus 112:**184–191
https://doi.org/10.1093/jhered/esab002

Hinrichs A. S. , Karolchik D. , Baertsch R. , Barber G. P. , Bejerano G. , Clawson H. , Diekhans M. , Furey T. S. , Harte R. A. , Hsu F. , Hillman-Jackson J. , Kuhn R. M. , Pedersen J. S. , Pohl A. , Raney B. J. , Rosenbloom K. R. , Siepel A. , Smith K. E. , Sugnet C. W. , Sultan-Qurraie A. , Thomas D. J. , Trumbower H. , Weber R. J. , Weirauch M. , Zweig A. S. , Haussler D. , Kent W. J. (2006) **The UCSC Genome Browser Database: update 2006** *Nucleic Acids Res* **34:**–598

Howe Kevin L , Achuthan Premanand , Allen James , Allen Jamie , Alvarez-Jarreta Jorge , Amode M Ridwan , Armean Irina M , Azov Andrey G , Bennett Ruth , Bhai Jyothish , Billis Konstantinos , Boddu Sanjay , Charkhchi Mehrnaz , Cummins Carla , Fioretto Luca Da Rin , Davidson Claire , Dodiya Kamalkumar , El Houdaigui Bilal , Fatima Reham , Gall Astrid , Giron Carlos Garcia , Grego Tiago , Guijarro-Clarke Cristina , Haggerty Leanne , Hemrom Anmol , Hourlier Thibaut , Izuogu Osagie G , Juettemann Thomas , Kaikala Vinay , Kay Mike , Lavidas Ilias , Le Tuan , Lemos Diana , Martinez Jose Gonzalez , Marugán José Carlos , Maurel Thomas , McMahon Aoife C , Mohanan Shamika , Moore Benjamin , Muffato Matthieu , Oheh Denye N , Paraschas Dimitrios , Parker Anne , Parton Andrew , Prosovetskaia Irina , Sak-thivel Manoj P , Salam Ahamed I Abdul , Schmitt Bianca M , Schuilenburg Helen , Sheppard Dan , Steed Emily , Szpak Michal , Szuba Marek , Taylor Kieron , Thormann Anja , Threadgold Glen , Walts Brandon , Winterbottom Andrea , Chakiachvili Marc , Chaubal Ameya , De Silva Nishadi , Flint Bethany , Frankish Adam , Hunt Sarah E , IIsley Garth R , Langridge Nick , Loveland Jane E , Martin Fergal J , Mudge Jonathan M , Morales Joanella , Perry Emily , Ruffier Magali , Tate John , Thybert David , Trevan-ion Stephen J , Cunningham Fiona , Yates Andrew D , Zerbino Daniel R , Flicek Paul (2020) **Ensembl 2021** *Nucleic Acids Research* **49:**
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

Huber Christian D. , Durvasula Arun , Hancock Angela M. , Lohmueller Kirk E. (2018) **Gene expression drives the evolution of dominance 9:**2750
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 19:**1195–1201

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

Korunes Katharine L. , Noor Mohamed A. F. (2017) **Gene conversion and linkage: effects on genome evolution and speciation 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 KirkE. (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 29:**197–205

Lewin Harris A. , Richards Stephen , Aiden Erez Lieberman , Allende Miguel L. , Archibald John M. , Bálint Miklós , Barker Katharine B. , Baumgartner Bridget , Belov Katherine , Bertorelle Giorgio , Blaxter Mark L. , Cai Jing , Caperello Nicolette D. , Carlson Keith , Castilla-Rubio Juan Carlos , Chaw Shu-Miaw , Chen Lei , Childers Anna K. , Coddington Jonathan A. , Conde Dalia A. , Corominas Montserrat , Crandall Keith A. , Crawford Andrew J. , DiPalma Federica , Durbin Richard , Ebenezer ThankGod E. , Edwards Scott V. , Fedrigo Olivier , Flicek Paul , Formenti Giulio , Gibbs Richard A. , Gilbert M. Thomas P. , Goldstein Melissa M. , Graves Jennifer Marshall , Greely Henry T. , Grigoriev Igor V. , Hackett Kevin J. , Hall Neil , Haussler David , Helgen Kristofer M. , Hogg Carolyn J. , Isobe Sachiko , Jakobsen Kjetill Sigurd , Janke Axel , Jarvis Erich D. , Johnson Warren E. , Jones Steven J. M. , Karlsson Elinor K. , Kersey Paul J. , Kim Jin-Hyoung , John Kress W. , Kuraku Shigehiro , Lawniczak Mara K. N. , Leebens-Mack James H. , Li Xueyan , Lindblad-Toh Kerstin , Liu Xin , Lopez Jose V. , Marques-Bonet Tomas , Mazard Sophie , Mazet Jonna A. K. , Mazzoni Camila J. , Myers Eugene W. , O'Neill Rachel J. , Paez Sadye , Park Hyun , Robinson Gene E. , Roquet Cristina , Ryder Oliver A. , Sabir Jamal S. M. , Bradley Shaffer H. , Shank Timothy M. , Sherkow Jacob S. , Soltis Pamela S. , Tang Boping , Tedersoo Leho , Uliano-Silva Marcela , Wang Kun , Wei Xiaofeng , Wetzer Regina , Wilson Julia L. , Xu Xun , Yang Huanming , Yoder Anne D. , Zhang Guojie (2022) **The earth biogenome project 2020: Starting the clock 119:** https://doi.org/10.1073/pnas.2115635118

Liu X. , Fu Y. X. (2015) **Corrigendum: Exploring population size changes using SNP frequency spectra 47:**1099

Ma Li , O'Connell Jeffrey R. , VanRaden Paul M. , Shen Botong , Padhi Abinash , Sun Chuanyu , Bickhart Derek M. , Cole John B. , Null Daniel J. , Liu George E. , Da Yang , Wiggans George R. (2015) **Cattle sex-specific recombination and genetic control from a large pedigree analysis 11:**1–24 https://doi.org/10.1371/journal.pgen.1005387

Ma Xin , Kelley Joanna L. , Eilertson Kirsten , Musharoff Shaila , Degenhardt Jeremiah D. , Martins André L. , Vinar Tomas , Kosiol Carolin , Siepel Adam , Gutenkunst Ryan N. , Bustamante Carlos D. (2013) **Population genomic analysis reveals a rich speciation and demographic history of orang-utans (Pongo pygmaeus and Pongo abelii) 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 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 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** *Journal of Dairy Science*
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 5:**

Meuwissen T H E , Hayes B J , Goddard M E (2001) **Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps 157:**1819–1829
https://doi.org/10.1093/genetics/157.4.1819

Miles A. , Harding N. J. , Botta G. , Clarkson C. S. , Antao T. , Kozak K. , Schrider D. R. , Kern A. D. , Redmond S. , Sharakhov I. , Pearson R. D. , Bergey C. , Fontaine M. C. , Donnelly M. J. , Lawniczak M. K. N. , Kwiatkowski D. P. , Donnelly M. J. , Ayala D. , Besansky N. J. , Burt A. , Caputo B. , Della Torre A. , Fontaine M. C. , Godfray H. C. J. , Hahn M. W. , Kern A. D. , Kwiatkowski D. P. , Lawniczak M. K. N. , Midega J. , Neafsey D. E. , O'Loughlin S. , Pinto J. , Riehle M. M. , Sharakhov I. , Vernick K. D. , Weetman D. , Wilding C. S. , White B. J. , Troco A. D. , Pinto J. , Diabaté A. , O'Loughlin S. , Burt A. , Costantini C. , Rohatgi K. R. , Besansky N. J. , Elissa N. , Pinto J. , Coulibaly B. , Riehle M. M. , Vernick K. D. , Pinto J. , Dinis J. , Midega J. , Mbogo C. , Bejon P. , Wilding C. S. , Weetman D. , Mawejje H. D. , Donnelly M. J. , Weetman D. , Wilding C. S. , Donnelly M. J. , Stalker J. , Rockett K. , Drury E. , Mead D. , Jeffreys A. , Hubbart C. , Rowlands K. , Isaacs A. T. , Jyothi D. , Malangone C. , Vauterin P. , Jeffery B. , Wright I. , Hart L. , Kluczy?ski K. , Cornelius V. , MacInnis B. , Henrichs C. , Giacomantonio R. , Kwiatkowski D. P. , Cornelius V. , MacInnis B. , Henrichs C. , Giacomantonio R. , Kwiatkowski D. P. (2017) **Genetic diversity of the African malaria vector Anopheles gambiae 552:**96–100

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

Nachman Michael W. (2002) **Variation in recombination rate across the genome: Evidence and implications 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 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 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.2021.637017

Pombi March , Stump Aram D. , Torre Allesandra Della , Besansky Nora J. (2006) **Variation in recombination rate across the x chromosome of anopheles gambiae 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 107:**583–588
https://doi.org/10.1016/j.ajhg.2020.08.017

Rhie Arang , McCarthy Shane A. , Fedrigo Olivier , Damas Joana , Formenti Giulio , Koren Sergey , Uliano-Silva Marcela , Chow William , Fungtammasan Arkarachai , Kim Juwan , Lee Chul , Ko Byung June , Chaisson Mark , Gedman Gregory L. , Cantin Lindsey J. , Thibaud-Nissen Francoise , Haggerty Leanne , Bista Iliana , Smith Michelle , Haase Bettina , Mountcastle Jacquelyn , Winkler Sylke , Paez Sadye , Howard Jason , Vernes Sonja C. , Lama Tanya M. , Grutzner Frank , Warren Wesley C. , Balakrishnan Christopher N. , Burt Dave , George Julia M. , Biegler Matthew T. , Iorns David , Digby Andrew , Eason Daryl , Robertson Bruce , Edwards Taylor , Wilkinson Mark , Turner George , Meyer Axel , Kautt Andreas F. , Franchini Paolo , Detrich III H. William , Svardal Hannes , Wagner Maximilian , Naylor Gavin J. P. , Pippel Martin , Malinsky Milan , Mooney Mark , Simbirsky Maria , Hannigan Brett T. , Pesout Trevor , Houck Marlys , Mis-uraca Ann , Kingan Sarah B. , Hall Richard , Kronenberg Zev , Sović Ivan , Dunn Christopher , Ning Zemin , Hastie Alex , Lee Joyce , Selvaraj Siddarth , Green Richard E. , Putnam Nicholas H. , Gut Ivo , Ghurye Jay , Garrison Erik , Sims Ying , Collins Joanna , Pelan Sarah , Torrance James , Tracey Alan , Wood Jonathan , Dagnew Robel E. , Guan Dengfeng , London Sarah E. , Clayton David F. , Mello Claudio V. , Friedrich Samantha R. , Lovell Peter V. , Osipova Ekaterina , Al-Ajli Farooq O. , Secomandi Simona , Kim Heebal , The-ofanopoulou Constantina , Hiller Michael , Zhou Yang , Harris Robert S. , Makova Kateryna D. , Medvedev Paul , Hoffman Jinna , Masterson Patrick , Clark Karen , Martin Fergal , Howe Kevin , Flicek Brian P. , Walenz Paul , Kwak Woori , Clawson Hiram , Diekhans Mark , Nassar Luis , Paten Benedict , Kraus Robert H. S. , Crawford Andrew J. , Gilbert M. Thomas P. , Zhang Guojie , Venkatesh Byrappa , Murphy Robert W. , Koepfli Klaus-Peter , Shapiro Beth , Johnson Warren E. , Di Palma Federica , Marques-Bonet Tomas , Teeling Emma C. , Warnow Tandy , Graves Jennifer Marshall , Ryder Oliver A. , Haussler David , O'Brien Stephen J. , Korlach Jonas , Lewin Harris A. , Howe Kerstin , Myers Eugene W. , Durbin Richard , Phillippy Adam M. , Jarvis Erich D. (2021) **Towards complete and error-free genome assemblies of all vertebrate species 592:**737–746
https://doi.org/10.1038/s41586-021-03451-0

Rosen Benjamin D , Bickhart Derek M , Schnabel Robert D , Koren Sergey , Elsik Christine G , Tseng Elizabeth , Rowan Troy N , Low Wai Y , Zimin Aleksey , Couldrey Christine , Hall Richard , Li Wenli , Rhie Arang , Ghurye Jay , McKay Stephanie D , Thibaud-Nissen Françoise , Hoffman Jinna , Murdoch Brenda M , Snelling Warren M , McDaneld Tara G , Hammond John A , Schwartz John C , Nandolo Wilson , Hagen Darren E , Dreischer Christian , Schultheiss Sebastian J , Schroeder Steven G , Phillippy Adam M , Cole John B , Van Tassell Curtis P , Liu George , Smith Timothy P L , Medrano Juan F (2020) **De novo assembly of the cattle reference genome with single-molecule sequencing 9:**
https://doi.org/10.1093/gigascience/giaa021

Schiffels Stephan , Wang Ke (2020) **MSMC and MSMC2: The Multiple Sequentially MarkovianCoalescent** 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 216:**499–519

Schrider Daniel R. , Kern Andrew D (2018) **Supervised Machine Learning for Population Genetics: A New Paradigm 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 8:**

Supek Fran , Lehner Ben (2019) **Scales and mechanisms of somatic mutation rate variation across the human genome** *DNA Repair* **81:**102647
https://doi.org/10.1016/j.dnarep.2019.102647

Talenti A , Powell J , Hemmink J D , Cook E A J , Wragg D , Jayaraman S , Paxton E , Ezeasor C , Obishakin E T , Agusi E R , Tijjani A , Marshall K , Fisch A , Ferreira B R , Qasim A , Chaudhry U , Wiener P , Toye P , Morrison L J , Connelley T , Prendergast J G D (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 118:**1–10
https://doi.org/10.1073/pnas.2015096118

Tennessen J. A. , Bigham A. W. , O'Connor T. D. , Fu W. , Kenny E. E. , Gravel S. , McGee S. , Do R. , Liu X. , Jun G. , Kang H. M. , Jordan D. , Leal S. M. , Gabriel S. , Rieder M. J. , Abecasis G. , Altshuler D. , Nickerson D. A. , Boerwinkle E. , Sunyaev S. , Bustamante C. D. , Bamshad M. J. , Akey J. M. (2012) **Evolution and functional impact of rare coding variation from deep sequencing of human exomes 337:**64–69

Teshima Kosuke M. , Coop Graham , Przeworski Molly (2006) **How reliable are empirical genomic scans for selective sweeps? 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 3:**711–721

VanRaden P M (2020) **Symposium review: How to implement genomic selection 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 1:**183–186

Wijnker E. , Velikkakam James G. , Ding J. , Becker F. , Klasen J. R. , Rawat V. , Rowan B. A. , de Jong D. F. , de Snoo C. B. , Zapata L. , Huettel B. , de Jong H. , Ossowski S. , Weigel D. , Koornneef M. , Keurentjes J. J. , Schneeberger K. (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 143:**941–952

## Author information


1. **M. Elise Lauterbur**

   Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85719, USA
   ORCID iD: 0000-0002-7362-3618

2. **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

3. **Ariella L. Gladstein**

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

4. **Graham Gower**

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

5. **Nathaniel S. Pope**

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

6. **Georgia Tsambos**

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

7. **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

8. **Saurabh Belsare**

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

9. **Arjun Biddanda**

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

10. **Victoria Caudill**

    Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA

11. **Jean Cury**

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

12. **Ignacio Echevarria**

School of Life Sciences, University of Glasgow, Glasgow, UK

13. **Benjamin C. Haller**

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

14. **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

15. **Xin Huang**

Department of Evolutionary Anthropology, University of Vienna, Vienna, Austria, Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Vienna, Austria

16. **Leonardo Nicola Martin Iasi**

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

17. **Ekaterina Noskova**

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

18. **Jana Obšteter**

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

19. **Vitor Antonio Corrêa Pavinato**

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

20. **Alice Pearson**

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

21. **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

22. **Manolo F. Perez**

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

23. **Murillo F. Rodrigues**

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

24. **Chris C. R. Smith**

Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402, USA

25. **Jeffrey P. Spence**

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

26. **Anastasia Teterina**

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

27. **Silas Tittes**

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

28. **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

29. **Juan Manuel Vazquez**

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

30. **Ryan K. Waples**

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

31. **Anthony Wilder Wohns**

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

32. **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

33. **Franz Baumdicker**

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

34. **Reed A. Cartwright**

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

35. **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

36. **Ryan N. Gutenkunst**

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

37. **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

38. **Andrew D. Kern**

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

39. **Aaron P. Ragsdale**

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

40. **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

41. **Daniel R. Schrider**

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

42. **Ilan Gronau**

Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel
ORCID iD: 0000-0001-8536-4062

## Editors

Reviewing Editor
**Ziyue Gao**
University of Pennsylvania

Senior Editor
**Molly Przeworski**
Columbia University, 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.

## 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.

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 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.

## 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.

## Author Response:

We are very glad that the 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 are preparing a revision of the preprint that will address these points.

One issue raised by the reviewers was that it is important to acknowledge possible limitations of the demographic model used in simulation in capturing different aspects of genomic variation. In particular, different demographic models inferred for the same species using different methods or sets of samples may have different strengths and weaknesses, and this should be considered when selecting a demographic model for simulation. This is an important point that we intend to discuss in the revised version of our manuscript. We also plan to expand the documentation of the stdpopsim catalog to include more information about the type of data used to fit every demographic model. Below we provide an outline of our thoughts on the topic.

First of all, it is important to acknowledge that demographic models inferred from genomic data cannot fully capture all aspects of the true demographic changes in the history of a species. As a result, these models do a good job in capturing some aspects of genetic variation, but not all of them. This is primarily determined by two factors: the method used for demographic inference, and the samples whose genomes were used in inference. Regardless of the method applied, the inferred demographic model can only reflect the genealogical ancestry of the sampled individuals, and this will typically make up a small portion of the complete genealogical ancestry of the species (albeit the genealogy of any set of sampled individuals includes many ancestors). Thus, demographic models inferred from larger sets of samples from diverse ancestry backgrounds may provide a more

comprehensive depiction of genetic variation within a species, as long as a sufficiently realistic demographic model can be fit. 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 behind PSMC-style inference methods).

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., dadi, Stairway plot), while other methods use haplotype structure (e.g., PSMC, MSMC, IBDNe). 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)). There have been several studies that compare different approaches to demography inference (e.g., Biechman et al. (2017 (https://academic.oup.com/g3journal/article/7/11/3605/6027516) ); Harris and Nielsen (2013 (https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003521) )), but unfortunately, there is currently no succinct handbook that describes the relative strengths and weaknesses of different methods. Indeed, we hope that the standardized simulations provided by stdpopsim will facilitate systematic comparisons between methods, which will, in turn, provide valuable insights for researchers when selecting demographic models for simulation.

It is important to 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. Scientists contributing a new model to the catalog are required to write a brief summary, which is added to the documentation page of the catalog: https://popsim-consortium.github.io/stdpopsim-docs/ latest/catalog.html. This summary includes a graphical description of the model (such as the one shown for *Anopheles gambiae* in Fig. 2B of the paper), as well as a description of the data and method used for inference. We will mention this in the revised manuscript to help users of stdpopsim navigate through this resource.