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

M. Elise Lauterbur[1], Maria Izabel A. Cavassim[2,*], Ariella L. Gladstein[3,*], Graham Gower[4,*], Georgia Tsambos[5,*], Jeff Adrion[6,7], Arjun Biddanda[8], Saurabh Belsare[6], Victoria Caudill[6], Jean Cury[9], Ignacio Echevarria[10], Benjamin C. Haller[11], Ahmed Hasan[12,13], Xin Huang[14,15], Leonardo Nicola Martin Iasi[16], Jana Obšteter[17], Vitor Antonio Corrêa Pavinato[18], David Peede[19,20], Ekaterina Noskova[21], Alice Pearson[22,23], Manolo Perez[24], Murillo F. Rodrigues[6], Chris C. R. Smith[6], Jeff Spence[25], Anastasia Teterina[6], Silas Tittes[6], Per Unneberg[26], Juan Manuel Vasquez[27], Ryan Waples[28], Anthony Wilder Wohns[29], Yan Wong[30], Reed Cartwright[31], Aaron P. Ragsdale[32], Franz Baumdicker[33], Gregor Gorjanc[34], Ryan N. Gutenkunst[35], Jerome Kelleher[30], Andrew D. Kern[6], Peter L. Ralph[6,36], Daniel R. Schrider[37], and Ilan Gronau[38]

[*]These authors contributed equally to the paper.
[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85719
[2]Department of Ecology and Evolutionary Biology University of California, Los Angeles
[3]Embark Veterinary, Inc., Boston, MA 02111, USA
[4]Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Denmark
[5]School of Mathematics and Statistics, University of Melbourne, Australia
[6]Institute of Ecology and Evolution, University of Oregon, Eugene OR 97402
[7]AncestryDNA, San Francisco, CA, 94107, USA
[8]54Gene, Inc., Washington, DC 20005, USA
[9]Université Paris-Saclay, CNRS, INRIA, Laboratoire Interdisciplinaire des Sciences du Numérique, UMR 9015 Orsay, France
[10]School of Life Sciences, University of Glasgow
[11]Department of Computational Biology, Cornell University
[12]Department of Cell and Systems Biology, University of Toronto, Toronto ON
[13]Department of Biology, University of Toronto Mississauga, Mississauga ON
[14]Department of Evolutionary Anthropology, University of Vienna, Djerassiplatz 1, 1030 Vienna, Austria
[15]Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Austria
[16]Department of Evloutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[17]Agricultural Institute of Slovenia, Department of Animal Science, Hacquetova ulica 17, Ljubljana, Slovenia
[18]Entomology Dept., CFAES, The Ohio State University, Wooster, Ohio
[19]Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA
[20]Center for Computational Molecular Biology, Brown University, Providence, RI, USA
[21]Computer Technologies Laboratory, ITMO University, St Petersburg, Russia
[22]Department of Genetics, University of Cambridge, UK
[23]Department of Zoology, University of Cambridge, UK
[24]Department of Genetics and Evolution, Federal University of Sao Carlos, Sao Carlos 13565905, Brazil
[25]Department of Genetics, Stanford University School of Medicine, Stanford, CA, 94305
[26]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Husargatan 3, SE-752 37 Uppsala, Sweden
[27]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA
[28]Department of Biostatistics, University of Washington
[29]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[30]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, OX3 7LF, UK
[31]School of Life Sciences and The Biodesign Institute, Arizona State University, Tempe, AZ USA
[32]Integrative Biology, University of Wisconsin-Madison, Madison, Wisconsin
[33]Cluster of Excellence - Controlling Microbes to Fight Infections, Eberhard Karls Universität Tübingen, Tübingen,

Baden-Württemberg, Germany
[34]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh EH25 9RG, UK
[35]Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721
[36]Department of Mathematics, University of Oregon, Eugene OR 97402
[37]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599
[38]Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel

September 20, 2022

Abstract

Simulation is a key tool in population genetics for both methods development and empirical research, but producing simulations that recapitulate even the main features of genomic datasets remains a major obstacle. Today, more realistic simulations are possible thanks to large increases in the quantity of available data and the sophistication of inference and simulation software~~, but~~ . However, implementing these simulations can require substantial time and specialized knowledge. These challenges are especially pronounced for simulating less well-studied species, since it is not always clear what level of realism is sufficient to confidently answer a given question, or what information is required to produce simulations of that desired realism. `Stdpopsim` is a community-developed ~~tool~~ framework that seeks to lower this barrier by making it easy to simulate complex population genetic models using up-to-date information. The initial version of ~~the~~ `stdpopsim` , the species catalog contained information for ~~6~~ six species, most of which are well-characterized model organisms. Here, we report on updates made in the new release of `stdpopsim`(version 0.2). In particular, we describe the community-driven efforts to expand the catalog more broadly across the tree of life, which now contains 21 species, with 25 demographic models and 37 genetic maps. ~~The process of expanding the catalog to include more speciesthrough community engagement yielded many insights, and we report on lessons learned~~ Our experience through the community engagement involved in this process was that people are indeed keen to put in the time and effort to include their study species, but that simple, clear guidance is vital. Our intention with this paper is in part to provide another learning modality to meet that need, by reporting on the main lessons learned through this process for best practices in population genomic simulation. We ~~discuss the elements of a population genomic simulationmodel, including the required input data,~~ describe the input data required for generating a realistic simulation, suggest good practices for obtaining the relevant information, and discuss common pitfalls and major considerations~~, and describe how new species models can be integrated into~~ . We also introduce several major advances to the realism of `stdpopsim`'s simulation ability, including gene conversion and provision of species-specific genomic annotations. Together, these advances to `stdpopsim` will strengthen efforts to use and develop simulation-based population genomic inference methods, with particular advances for non-model organisms, making them available, transparent, and accessible to everyone.

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 (for example, Vertebrate Genomes (Rhie et al., 2021), 10,000 Plants (Cheng et al., 2018) and others (Darwin Tree of Life Project Consortium, 2022)) are providing the backbone for enormous increases in the amount of population-level genomic data available for model and non-model species. These data are being used to answer questions across scales from deep evolutionary time to ongoing ecological dynamics. Methods that use these data, for example to infer 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 population dynamics not initially accounted for, such as inbreeding or selfing (Blischak et al., 2020), skewed offspring distributions (Montano, 2016), and intense artificial selection (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 begun to be 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 detailed elements of a species' biology. Important elements include genomic features such as mutation and recombination rates that strongly affect genetic variation and haplotype structure (Nachman, 2002). These have particularly strong ramifications when linked selection is important in 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 (e.g., those governing the process of natural selection) are fundamentally important when developing 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.

`Stdpopsim` is a community resource recently developed to provide 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 ~~from workshops~~ we received from the community identified a widespread desire to simulate a wider range of non-model species, and ideally to incorporate these into the `stdpopsim` catalog for future use. ~~That~~ This feedback, and subsequent efforts to expand the catalog, also uncovered the need for a better understanding of when it is practical to create a realistic simulation of a species of interest, and indeed what "realistic" means in this context. ~~In addition to 's framework for standardizing simulations of some species, our experience has led us to develop guidance that may be of use to the broader population genetics community.~~

This paper ~~is intended to announce and describe the additions to the~~ reports on the updates made in the current release of `stdpopsim` ~~catalog, and~~ (version 0.2), and is also intended as a resource for ~~methods developers and empirical researchers who wish to develop simulations of~~ any researcher who wishes to develop whole-genome simulations for their own species of interest~~or add~~. We start by describing the main idea behind the standardized simulation framework of `stdpopsim`, and then outline the main updates made to the `stdpopsim` catalog ~~. In the section , we discuss the elements of a population genomic simulation model that characterizes a species, including~~ and simulation framework in the past two years. We then devote a major section of the paper to provide guidelines for generating population genomic simulations, either for the purpose of using them in one specific study, or with the intent of adding these simulations to `stdpopsim`. Among other things, we discuss when a whole-genome simulation is more useful than simulations based on either individual loci or generic (non-species specific) loci. We ~~discuss~~ specify the required input data~~(genome assembly, mutation and recombination rates, and demographic model),~~ , mention common pitfalls in choosing appropriate parameters, and ~~considerations~~ suggested courses of action for species that are missing estimates of some necessary inputs. ~~This paper is not intended as a tutorial for implementing simulationsin any particular simulator, rather to provide guidance for what information is sufficient for a realistic genome~~ We conclude with examples from a couple of species recently added to `stdpopsim`, which demonstrate some of the main considerations involved in the process of designing realistic whole-genome simulations. While the guidelines provided in this paper are intended for any researcher interested in implementing a population genomic simulation using any ~~simulator. We pay particular attention to~~ software, we do highlight the ways in which the framework set up by `stdpopsim` eases ~~this burden , and describe how new users might add their own species information to . The latter is discussed in the section, where we lay out in detail the simple~~

~~processof incorporating the information discussed in the section into~~ the burden involved in this process.

The utility of stdpopsim for genome-wide simulations

~~We begin with an~~ We begin by providing a brief overview of the ~~goals and~~ importance of genome-wide simulations and the main rationale behind stdpopsim ~~and complete chromosome simulation~~; 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 known conditions that model a given species (or population) of interest. 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, that 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. ~~As a result, while the simulations themselves may require considerable computational resources, the most time-consuming and~~ Thus, the process of coding a realistic simulation can be quite time consuming and often error-prone~~part of population genomic simulation is often the task of correctly parameterizing simulation software~~.

The main objective of stdpopsim is to streamline this process, ~~making it less time consuming, less error prone,~~ and to make it more robust and more reproducible. Contributors ~~use a template to build the model~~ collect parameter values for their species of interest ~~, including the required parameter values.~~ from the literature, and then specify these parameters in a template file for the new model. This model then goes through a vital peer-review process, ~~including validating the choices of parameter values. Any discrepancies are resolved~~ which involves recreating the model based on the provided documentation, and executing automated scripts 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 considerably~~ Importantly, this increases the reliability of the resulting simulations in any downstream analysis.

~~The goal of complete chromosome simulation is important for a number of reasons. The~~ Another central goal of stdpopsimis to promote whole-genome 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 important benefits. First, the organization of genes on chromosomes is a key feature of a species' genome ~~, and one that has largely been ignored in population genomic simulation~~ that is clearly ignored in traditional population genomic simulations (see Schrider (2020) for a notable exception). ~~This is largely because simulation of chromosome-scale sequences, on the order of $> 10^7$ bp, has until recently been largely out of reach computationally, so population geneticists have resorted to separate simulations of many short segments of the genome (e.g., Harris and Nielsen, 2016).~~

~~However, physical linkage of chromosomes induces correlations along a chromosome that generally~~ Second, modeling physical linkage allows simulations to capture important correlations between genetic variants along the same chromosomes. These correlations reduce variance relative to independent simulations of equivalent genetic material. This has a particularly striking effect in long stretches of low recombination rates, 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 gene conversion of short segments. 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 genome variation in myriad 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 ~~. Methods~~ on their demographic history. Demography inference methods that use such information, such as MSMC (Schiffels and Wang, 2020), or 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 for design of empirical studies that use them.

~~Additions to~~

Additions to `stdpopsim`

Since its initial publication in Adrion et al. (2020), we have increased the number of species in the catalog nearly fourfold, added multiple demographic models and genetic maps, and improved the simulation framework of `stdpopsim` in several ways.

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 introduction of additional demographic models for *Homo sapiens*, *Pongo abelii*, *Drosophila melanogaster*, and *Arabidopsis thaliana*, enabling a wider variety of simulations for these mostly model species.

However, these species represent a small slice of the tree of life. This is a concern not only because ~~of the~~ there is a large community of researchers studying other organisms~~that might benefit from these efforts~~, but also because methods developed for application to ~~humans (for instance~~model species (such as humans) may not perform well when applied to other species with very different biology. ~~It should thus be made easy to test~~ Adding species to the `stdpopsim`catalog will allow developers to easily test their methods across a ~~wide~~ wider variety of organisms. ~~To begin to address this, we~~

We thus made a concerted effort to recruit members of the population and evolutionary genetics community to add ~~new species~~ their species of interest to the `stdpopsim` catalog~~, culminating in~~. 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. ~~To introduce people to using and to prepare people for the hackathon, we organized a series of seven workshops in the preceeding months. These~~ The seven 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 for the hackathon, 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, which resulted in the addition of 15 species to the `stdpopsim` catalog (Figure 1).

~~Phylogenetic tree of species available in the catalog. In blue are species we published in the original release (Adrion et al., 2020), in orange are those species that have since been added. Columns show which species have one (light grey) or more (dark grey) demographic models and genetic maps.~~

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*), ~~and~~ two bacteria, ~~in addition to~~ four primates and a common mammalian associate of primates (*Canis familiaris*). Not all of these have genetic maps or demographic models (see Figure 1), but this lays the framework for future contributions.

~~A key feature added to that expands the diversity of species that can be realistically modeled is the inclusion of recombination by gene conversion, which is essential for organisms such as *E. coli* that lack crossing over. In addition, the update to msprime~~ Expanding the species catalog required adding several functionalities to the simulation framework. We thus upgraded the neutral simulation engine, `msprime`, from version 0.7.4 to version 1.0 ~~(Baumdicker et al., 2021)~~. (Baumdicker et al., 2021). This upgrade provides a number of ~~additional~~ benefits such as a discrete site model of mutation, so that simulated data will now ~~,~~

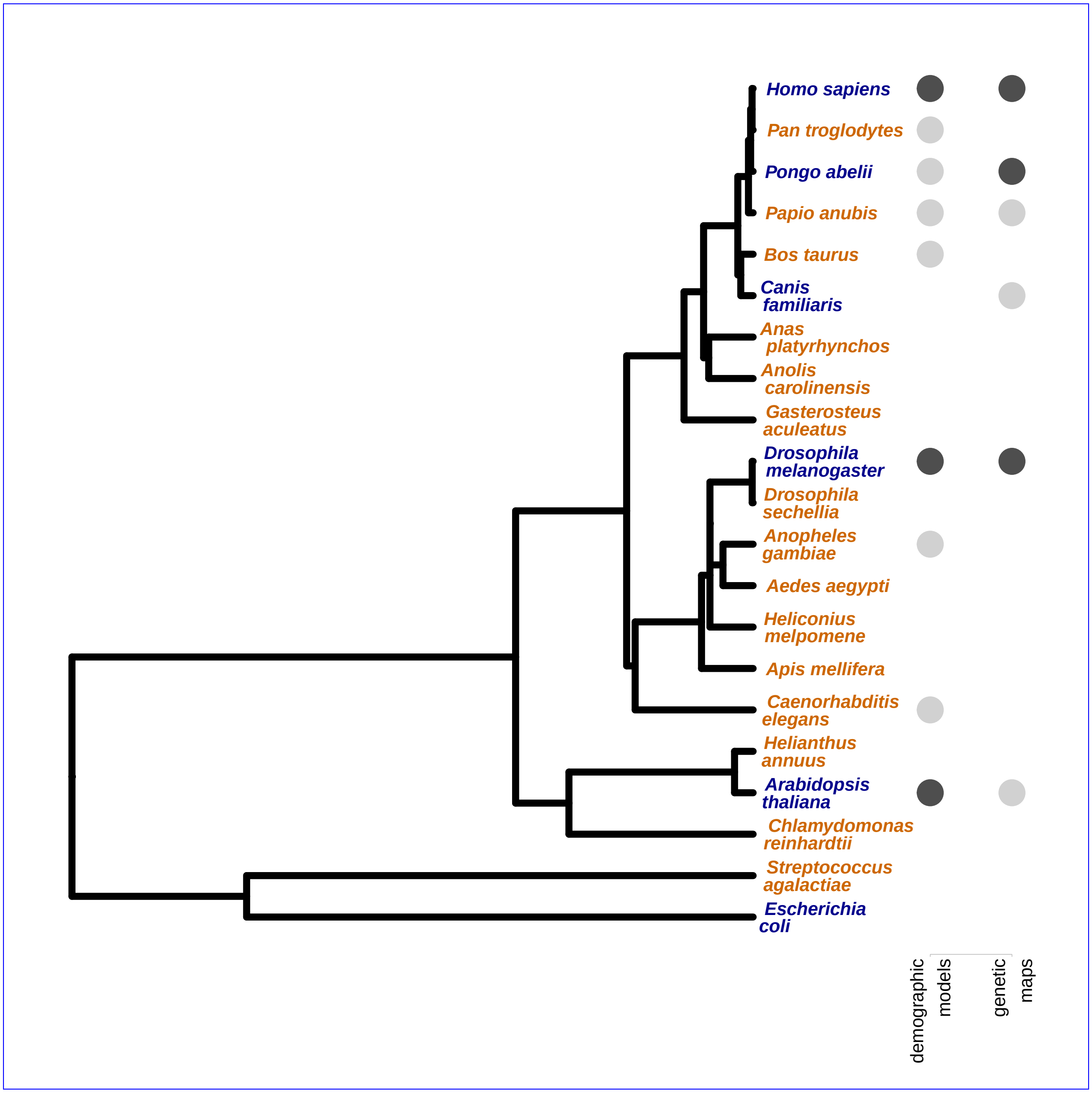

Figure 1: Phylogenetic tree of species available in the stdpopsimcatalog, including the six species we published in the original release (Adrion et al., 2020, in blue), and 15 species that have since been added (in orange). Solid circles indicate species that have one (light grey) or more (dark grey) demographic models and genetic maps.

~~as in real data,~~ have a small proportion of sites with multiple mutations and possibly more than two alleles. Another key feature enabled by this upgrade was recombination by gene conversion, which is essential for modeling genomes of bacteria and archea. Gene conversion affects shorter segments than crossover recombination and creates distinct patterns of genetic diversity along the genome (Korunes and Noor, 2017) . In bacteria and archaea, homologous recombination occurs primarily through gene conversion rather than single crossover recombinations. As a result, such species cannot be realistically simulated with a recombination model that considers only crossovers. Gene conversion in such species is implemented in stdpopsimby setting the bacterial_recombination flag to True (as a feature of the genome) and the gene_conversion_length parameter to the average gene conversion tract length. This will result in all recombinations being simulated via gene conversion (no crossover recombinations), where the tract length is sampled from a geometric distribution whose mean is the specified length. For example, the model for *Escherichia coli* has been updated in the stdpopsimcatalog to have a gene conversion rate of 8.9×10^{-11} with an average tract length of 345 bases (Wielgoss et al., 2011; Didelot et al., 2012). Some species undergo recombination by gene conversion as well as crossover recombination. To accommodate this in stdpopsimsimulations, one needs to set two additional parameters in each chromosome: gene_conversion_fraction, which specifies the fraction of recombinations that occur due to gene conversion; and gene_conversion_length, which is the average tract length, as defined above. For example, the model for *Drosophila melanogaster* has been updated in the stdpopsimcatalog to have a fraction of gene conversions of 0.83 (in all chromosomes with recombination) and an average tract length of 518 bases (Comeron et al., 2012).

~~Moreover, we have extended~~ Lastly, we extended stdpopsim ~~so that genome annotations can be associated with~~ to allow augmenting a genome assembly ~~. These~~ by genome annotations, such as coding regions, promoters, conserved elements, etc. These annotations can be used to simulate selection at a subset of sites (e.g., the annotated coding regions) using parametric distribution(s) of fitness effects. ~~This step is transformative – standardized~~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)~~, and through~~. Thus, we expect this extension of stdpopsim ~~this is now achievable. However, this is not the focus of the current paper, since these~~ to be transformative in the way simulations are carried out in population genetics. These significant new capabilities of the stdpopsim library will be detailed in a forthcoming publication, and are not the focus of this paper.

Guidelines for implementing a population genomic simulation

The concentrated effort to add species to the stdpopsim catalog has lead to a series of important insights about this process, which we summarize in the following section as a set of guidelines for implementing realistic simulations of any species. ~~We stress that such an implementation could be done using any software or engine~~Our intention is to provide general guidance that applies to any population genomic simulation software, but we ~~here pay special attention to the case with which this can be accomplished~~ also mention specific requirements that apply to simulations done in the framework of stdpopsim.

Basic setup for chromosome-level simulations

Implementing a realistic population genomic simulation for a species of interest requires ~~integrating information from several publications to choose appropriate parameter values. In this section, we outline these pieces of information and provide guidelines for how to use them to set the simulation parameters.~~

~~Basic setup for chromosome-level simulations~~

~~To run a simulation requires a~~ a fairly detailed description of the organism's demography and mechanisms of genetic inheritance. ~~Although in practice many parameters describing these processes might be only roughly guessed at,~~ While simulation software requires unforgivingly precise values~~. We start by describing how and where to find appropriate values, and some possible alternatives when values for the ideal parameters are not known.~~ , in practice, we may only have rough guesses for most of the parameters describing these processes. In this section, we list these 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. Currently, the number of species with complete chromosome-level assemblies is small, but we expect this number to dramatically increase in the near future due to genome initiatives such as the Earth Biogenome (Lewin et al., 2022) and its affiliated project networks (e.g., Vertebrate Genomes (Rhie et al., 2021), 10,000 Plants (Cheng et al., 2018)). Furthermore, the development of new long-read sequencing technologies (Amarasinghe et al., 2020, 2021) and concomitant advances in assembly pipelines (Chakraborty et al., 2016) are likely to boost these initiatives. When expanding the `stdpopsim` catalog, we decided to focus on species with near-complete chromosome-level genome assemblies (i.e., close to one contig per chromosome). This restriction was set mainly because species with less complete genome builds typically do not have good estimates of recombination rate or genetic maps, making chromosome-level simulation much less useful. Therefore, the utility of adding such species to the catalog does not justify the maintenance and storage burden incurred by the large number of contigs in these partial assemblies (see also discussion below).

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). ~~Although mutation rates Benzer1961~~At present, ~~Ellegren2003and processes Supek2019are not uniform along the genome or through time, at presentmutations are simulated~~ `stdpopsim`simulates mutations at a constant rate under the Jukes-Cantor model of nucleotide mutations ~~(CITE). We antcipate future efforts~~ (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~~ `SLiM` and `msprime` simulation engines. Such progress will further improve the realism of simulated genomes, since mutation rates and processes are known to vary along the genome and through time (Benzer, 1961; Ellegren et al., 2003; Supek and Leh
.

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 ~~Nachman2002~~(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 genome-wide recombination rate needs to be specified, which is typically available for well assembled genomes. Recall that for bacteria and archea, which primarily experience recombination by gene conversion, the recombination rate corresponds to the rate of gene conversion, and the average tract length should also be specified (see details in previous section). If one wishes to model gene conversion together with crossover recombination, then they should specify the fraction of recombinations done by gene conversion as well as the average tract length (per chromosome).

4. **A demographic model** ~~describing the history of the population, e. g., by specifying historical~~ 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. ~~Since misspecification of the demographic modelcan generate unrealistic patterns of genetic variation that will affect downstream analyses (e.g., Navascués and Emerson, 2009).~~ 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 their study. 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 reproduce in simulation the average observed genetic diversity in that species. Note, however, that this average effective population size will not capture features of genetic variation that are caused by recent changes in population size and the presence of population structure ~~(MacLeod et al., 2013).~~ (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 popu-

lation 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~~`SLiM`) or the Moran model (in ~~msprime~~`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. As mentioned above, ~~the~~ it is a widely appreciated fact that linked selection modulates patterns of variation within genomes~~necessitates its inclusion~~ . Therefore, its incorporation into simulations is crucial for many purposes. ~~For~~ 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. ~~This release~~ The current version of `stdpopsim` ~~includes a way to describe these features, and the ability to simulate selection on these regions~~ enables simulation with selection (using the ~~SLiM engine)~~ `SLiM` engine) by specifying genome annotations and distributions of fitness effects, as specified below. We note that the ability to simulate chromosomes with realistic models of selection is still under development and will be finalized in the next release of `stdpopsim` .

6. **Genome annotations**, specifying regions subject to selection (e.g., as 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 neutrally evolving.

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 desireable to choose parameters that best reflect our current understanding. 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 from word of mouth. However, values in `stdpopsim` are always chosen to match published estimates, so that the underlying data and methods are documented ~~. Another key practice within is quality control:~~ 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, some kind of 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. ~~This practice often finds subtle bugs and helps increase the reliability and reproducibility of the catalog. We highly recommend the similar practice of code review for simulations generated outside of .~~

Obtaining reliable and citeable 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 some care, because 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 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 ~~have~~ be simulated using a mutation rate that differs from the default rate specified for the species~~, which will be used when the model is simulated.~~

~~This~~ . See, for example, the model implemented for *Bos taurus*, which is described in the next section. This important feature does not necessarily fix all ~~inconsistencies, due to other~~ potential inconsistencies caused by assumptions made by the demographic inference method ~~that are not captured by the simulation, such as assuming a recombination rate different than the one we use for the species model~~(such as assumptions on recombination rates). It is therefore ~~simpler~~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 inconsistiences~~ for inconsistency is coordinate drift between ~~current reference genomes assembliesand previously constructed annotations or genetic maps. Following~~ subsequent versions of genome assemblies. In `stdpopsim`, we follow the approach from the UCSC Genome Browser ~~, in we~~ and use liftover to ~~align~~ convert the coordinates of ~~the genetic maps~~ genetic maps and genome annotations that we curate to the coordinates of the ~~reference genome assemblies~~genome assembly we use for that species.

Filling out the missing pieces

For many species it is difficult to obtain estimates of ~~the~~ all necessary model parameters. ~~We provide several suggestions for dealing with this scenario (see Table 1).~~ Table 1 provide suggestions for ways to deal with missing values of various central model parameters. The table also mentions the main discrepancies between the simulated data and real genomic data, which can be caused by mis-specification of each parameter.

Several researchers who participated in ~~our hackathon in 2020~~ the "Growing the Zoo" hackathon wished to add species whose genome assemblies are composed of many relatively small contigs, unanchored to chromosome-level scaffolds. Although ~~previously we did not plan to have restrictions on which species might be added, we decided that we would~~ we wish to keep `stdpopsim`as inclusive as possible, we made a conscience decision to only add species with chromosome-level assemblies. ~~One consideration behind this decision is load time for the library: species with tens of thousands of contigs require these lists of contig lengths (and associated information) to be loaded at runtime. However, the same issue exists for genetic maps, which is why these do not come pre-loaded but are downloaded from cloud storage upon first use. The second consideration is that the purpose of is to make complex simulations easy, i.e., to streamline the loading in of complex information that will make the simulation more realistic, such as genetic maps~~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~~. However, species with fragmentary assemblies generally do not have estimates of complex demographic models, nor genetic maps. Finally~~, ~~although we could crowd-source addition of many species, still each one required substantial attention by a core group of maintainers~~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 after changes to the `stdpopsim`framework. So, the benefit of including ~~such species~~ species with very partial genome builds in `stdpopsim` would be outweighed by the substantial extra burden

Table 1: **Guidelines for dealing with missing parameters.** For each parameter, we provide a suggested course of action, and mention the main discrepancies between the simulated data and real genomic data, which can be caused by mis-specification of that parameter.

Missing parameter	**Suggested action**	**Possible discrepancies**
Mutation rate	Borrow from closest relative with a citeable mutation rate	Number of polymorphic sites
Recombination rate	Borrow from closest relative with a citeable recombination rate	Patterns of linkage disequilibrium
Gene conversion rate and tract length	Set rate to 0 or borrow from closest relative with a citeable rate	Lengths of shared haplotypes across individuals
Demographic model	Set the effective population size (Ne) 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

on ~~downstream users and~~ `stdpopsim` maintainers as well as downstream users of these models.

~~However~~That being said, simulation is still ~~useful in such species~~ possible and potentially useful for species with partial genome builds. One way to deal with this situation is to include only the longer contigs or scaffolds, treating them as separate chromosomes in the simulation. Some of these contigs will map to the same chromosome, so simulating them separately will not capture the genetic linkage between them. However, this provides a reasonable approximation for many purposes, at least for genomic regions far from the contig edges. Short contigs can either be omitted from simulation, or lumped together into one (or several) longer pseudo-chromosome(s). ~~We caution that this has the potential to result in false precision when these effects are present in the real genome but missing from the diversity generated by the simulation. Finally, although whole-chromosome simulations are crucial for many purposes, for~~ Creating pseudo-chromosomes allows the simulation to fit the amount of data of real genomes, but it artificially increases the correlation between variants. Finally, we note that for some situations it may be sufficient to rely on simulation of ~~many~~ 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, we caution that in general the influence of linkage on the uncertainty of such inferences is not well understood. An alternative is to instead simulate an anonymous chromosome from which patterns of genetic variation can be extracted (if important, in chunks of size similar to the contigs). The latter is usually more realistic, since this includes linkage between sites that share a chromosome but may be on different real contigs. Precise locations in the simulated genomes cannot then be matched to particular contigs, but general statistical patterns can be compared.~~ However, this approach would not have the many benefits of whole-chromosome simulations, which we discussed in detail earlier.

~~Missing parameter Options Considerations Mutation rate borrow from closest relative with a citeable mutation rate will affect levels of polymorphism Recombination rate borrow from closest relative with a citeable rate will affect the impact of selection, linkage, and linked selection Demographic model at least Ne is required and is estimable from mutation rate and genetic data the demographic history (e.g. bottlenecks, expansions, and population splits and migration) affects patterns of variation substantially CITE, a constant Ne is not ideal~~

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 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 ~~simulation are~~ species model is based on the AgamP4 **genome assembly** (Sharakhova et al., 2007), which was downloaded from Ensembl (Howe et al., 2020) via `stdpopsim`'s utilities that interact with Ensembl. These utilities make it easy to accurately retrieve basic genome information and construct the appropriate Python data structures.

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

[FIG TBA]

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.

$\mu = 3.5 \times 10^{-9}$ mutations per generation per site, estimated for *D. melanogaster* by Keightley et al. (2009) and 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 ~~this mutation rate , the~~ the formula $\theta = 4\mu N_e$, with the above mutation rate ($\mu = 3.5 \times 10^{-9}$), and a mean nucleotide diversity of ~~the samples from Gabon reported in Miles et al. (2017) , and the relation $\theta = 4\mu N_e$. This results~~ $\theta \approx 0.015$, as reported by Miles et al. (2017) for the Gabon population. This resulted in an estimate of ~~N_e close to 10^6~~$N_e = 1.07 \times 10^6$, which we rounded down to one million. These steps were documented in the code for the `stdpopsim` species model~~. In doing this we made some arbitrary choices: which sampling location to use data from, and how to round the resulting estimate. However, these choices were not worrisome, since a single~~ , 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 ~~provides only a very rough approximiation to the demographic history of samples from any region. 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).~~ is meant to provide only a rough approximation to historic population sizes, which is to be overwritten by a more detailed demographic model.

Miles et al. (2017) inferred **demographic models** from *Anopheles* samples from ~~9 locations~~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, ~~a model~~ which consists of a single population whose size changes throughout the past 11,260 generations in 67 time intervals ~~.~~ (Figure 2B) During this time period, the population size was inferred to have fluctuated from below 80,000 (an ancient bottleneck roughly 10,000 generations ago) to the present-day estimate of over 4 million individuals. To convert the timescale from generations to years, we used an average generation time of 1/11 years, as in Miles et al. (2017).

All of these parameters were set in the appropriate source files in the `stdpopsim` catalog, accompanied by the relevant citation ~~infromation. The species~~ information, and the model underwent the standard quality control process~~before it was added to the catalog. It~~ . The model may be refined in the future by adding more demographic models~~or updating the mutation rate estimate or~~ , updating or refining the recombination map, or updating the mutation rate estimates based on ones directly estimated for this species. Note that ~~if in the future we obtain a direct estimate of mutation rate for *Anopheles gambiae*, then~~ even if the mutation rate is ever updated, the demographic model mentioned above should ~~be appropriately rescaled to match the new mutation rate~~ 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 increasingly 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). ~~Ancestral~~ 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 will not generate realistic genome structure and diversity (MacLeod et al., 2013; Rosen et al., 2020). ~~However, the software~~ Since `stdpopsim` does not allow for a missing value of N_e ~~(and we chose not to change this requirement), so~~ , we chose to set the species default N_e to the ancestral estimate of 6.2×10^4~~, but~~ . 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 the reduction from the ancestral effective population size ($N_e = 62,000$) beginning around 33,000 generations ago, consisting of a series of 13 instantaneous population size changes, ultimately reaching the current effective population size ($N_e = 90$) in the 1980s (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 (1.2×10^{-8} from Harland et al., 2017, as used above) is thus 28% higher than the rate used for inference. As a result, if one were to simulate the demographic model with the species' default mutation rate, they would produce synthetic genomes with considerably higher sequence diversity than actually observed in real genomic data. To address this, we specified a mutation rate of $\mu = 9.4 \times 10^{-9}$ in the demographic model, which then overrides the species' mutation rate when this demographic model is applied in simulation. The issue of fitting the rates used in simulation with those assumed during inference was discussed during the independent review of this demographic model, and it raised an important question about recombination rates. Since MacLeod et al. (2013) use runs of homozygosity to infer the demographic model, their results depends on the assumed recombination rate. The recombination rate assumed in inference ($r = 10^{-8}$) 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 simulated genomes 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 modeled here.

Conclusion

As our ability to sequence genomes continues to advance, the need for population genomic simulation of new model and non-model organism genomes is becoming acute. So too is the concomitant need for an expandable framework for implementing such simulations for species of interest and the resources for understanding when and how to do so.

Simulating species of interest, both model and non-model, presents significant challenges in coding and the choice of parameter values on which to base the simulation. `Stdpopsim` is a resource that is uniquely poised to address these challenges as it provides easy access to simulations incorporating species-specific information, easy inclusion of new species genomes, and the choices of new species to include are driven by the needs of the population genomics community. In this manuscript we describe the expansion of `stdpopsim` in two ways: the expansion of its underlying framework to incorporate new evolutionary processes such as gene conversion, which broadens 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 present basic considerations for implementing population genomic simulations, agnostic to simulation software, based on insights from the community-driven process of expanding the `stdpopsim` catalog. We describe the steps of determining if a species-specific population genomic simulation is appropriate for the species and question, what data is necessary and why, special considerations for finding and using that data, how to proceed when some of that data is not available, and why we encourage everyone implementing simulations to have their parameter choices and implementation reviewed by at least one other researcher. These steps can be followed independently, or, as we encourage, through the stdpopsim framework for quality control and to make the species model available for future standardized 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, with some prior knowledge of mutation and recombination rates, will become a candidate for inclusion into the stdpopsim catalog following the steps we have outlined above. As annotations of those genome assemblies improve over time this information too can easily be added to the stdpopsim catalog.

Moreover, one of the goals of stdpopsim is to leverage stdpopsim itself as a springboard for education and inclusion of new communities into computational biology and software development. We are keen to use outreach, for instance in the form of workshops and hackathons described here, as a way to democratize development of population genomic simulation as well as grow the stdpopsim catalog and library generally. By enabling researchers of non-model species with simulation platforms that traditionally have been quite narrowly focused with respect to organism, we hope to improve the ease and reproducibility of research across a large number of systems, while simultaneously expanding the community of software developers at work in the population and evolutionary genetics world. Our experience with such outreach over the past two years is that people are indeed keen to put in the time and effort to include their study species, but that simple, clear guidance is vital. Our intention with this paper is in part to provide another learning modality to meet that need.

Acknowledgements

~~TODO Workshop and hackhaton attendes?~~ We wish to thank the dozens of workshop attendees, and especially to the two dozen or so hackathon members, whose combined feedback motivated many of the updates made to stdpopsimin the past two years.

Funding

~~TODO: should we order these alphabetically or in the same order as authors?~~

M. Elise Lauterbur was supported by an NSF Postdoctoral Research Fellowship #2010884. 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. Andrew D. Kern and Peter L. Ralph were supported by NIH award R01HG010774~~. Ryan N. Gutenkunst was supported by NIH award R01GM127348~~.

References

Jeffrey R Adrion, Christopher B Cole, Noah Dukler, Jared G Galloway, Ariella L Gladstein, Graham Gower, Christopher C Kyriazis, Aaron P Ragsdale, Georgia Tsambos, Franz Baumdicker, Jedidiah Carlson, Reed A Cartwright, Arun Durvasula, Ilan Gronau, Bernard Y Kim, Patrick McKenzie, Philipp W Messer, Ekaterina Noskova, Diego Ortega-Del Vecchyo, Fernando Racimo, Travis J Struck, Simon Gravel, Ryan N Gutenkunst, Kirk E Lohmueller, Peter L Ralph, Daniel R Schrider, Adam Siepel, Jerome Kelleher, and Andrew D Kern. A community-maintained standard library of population genetic models. *eLife*, 9:e54967, jun 2020. ISSN 2050-084X. doi: 10.7554/eLife.54967. URL https://doi.org/10.7554/eLife.54967.

Shanika L. Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21, 2020. doi: https://doi.org/10.1186/s13059-020-1935-5.

Shanika L Amarasinghe, Matthew E Ritchie, and Quentin Gouil. long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data. *GigaScience*, 10(2), 02 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab003. URL `https://doi.org/10.1093/gigascience/giab003`. giab003.

Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, Ariella L Gladstein, Gregor Gorjanc, Bing Guo, Ben Jeffery, Warren W Kretzschumar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S Pope, Consuelo D Quinto-Cortés, Murillo F Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W Wohns, Yan Wong, Simon Gravel, Andrew D Kern, Jere Koskela, Peter L Ralph, and Jerome Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3), 12 2021. ISSN 1943-2631. doi: 10.1093/genetics/iyab229. URL `https://doi.org/10.1093/genetics/iyab229`. iyab229.

Annabel C. Beichman, Emilia Huerta-Sanchez, and Kirk E. Lohmueller. Using genomic data to infer historic population dynamics of nonmodel organisms. *Annu. Rev. Ecol. Evol. Syst.*, 49:433–456, 2018. ISSN 15452069. doi: 10.1146/annurev-ecolsys-110617-062431.

Seymour Benzer. On the topography of the genetic fine structure. *Proceedings of the National Academy of Sciences*, 47(3):403–415, 1961. doi: 10.1073/pnas.47.3.403. URL `https://www.pnas.org/doi/abs/10.1073/pnas.47.3.403`.

Paul D. Blischak, Michael S. Barker, Ryan N. Gutenkunst, and Daniel Falush. Inferring the Demographic History of Inbred Species from Genome-Wide SNP Frequency Data. *Mol. Biol. Evol.*, 37(7):2124–2136, 2020. ISSN 15371719. doi: 10.1093/molbev/msaa042.

Sharon R. Browning and Brian L. Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*, 97 (3):404–418, 2015. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2015.07.012. URL `https://www.sciencedirect.com/science/article/pii/S0002929715002888`.

A Cesarani, D Lourenco, S Tsuruta, A Legarra, E L Nicolazzi, P M VanRaden, and I Misztal. Multibreed genomic evaluation for production traits of dairy cattle in the United States using single-step genomic best linear unbiased predictor. *Journal of Dairy Science*, 105(6):5141–5152, 2022. doi: https://doi.org/10.3168/jds.2021-21505.

Mahul Chakraborty, James G Baldwin-Brown, Anthony D Long, and JJ Emerson. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic acids research*, 44(19): e147–e147, 2016.

B. Charlesworth. The effects of deleterious mutations on evolution at linked sites. *Genetics*, 190(1):5–22, Jan 2012.

Shifeng Cheng, Michael Melkonian, Stephen A. Smith, Samuel Brockington, John M. Archibald, Pierre-Marc Delaux, Fay-Wei Li, Barbara Melkonian, Evgeny V. Mavrodiev, Wenjing Sun, Yuan Fu, Huanming Yang, Douglas E. Soltis, Sean W. Graham, Pamela S. Soltis, Xin Liu, Xun Xu, and Gane Ka-Shu Wong. 10kp: A phylodiverse genome sequencing plan. *Gigascience*, 3(7), 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy013.

Chris S Clarkson, Alistair Miles, Nicholas J Harding, Eric R Lucas, CJ Battey, Jorge Edouardo Amaya-Romero, Andrew D Kern, Michael C Fontaine, Martin J Donnelly, Mara KN Lawniczak, et al. Genome variation and population structure among 1142 mosquitoes of the african malaria vector species anopheles gambiae and anopheles coluzzii. *Genome research*, 30(10):1533–1546, 2020.

J. M. Comeron, R. Ratnappan, and S. Bailin. The many landscapes of recombination in Drosophila melanogaster. *PLoS Genet*, 8(10):e1002905, 2012.

Josep M Comeron. Background selection as baseline for nucleotide variation across the drosophila genome. *PLoS Genetics*, 10(6):e1004434, 2014.

Katalin Csilléry, Michael G B Blum, Oscar E Gaggiotti, and Olivier François. Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.*, 25(7):410–8, jul 2010. ISSN 0169-5347. doi: 10.1016/j.tree.2010.04.001. URL `http://www.ncbi.nlm.nih.gov/pubmed/20488578`.

A. D. Cutter and B. A. Payseur. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetcs*, 14(4):262–274, 2013. doi: https://doi.org/10.1038/nrg3425. URL `https://www.nature.com/articles/nrg3425`.

Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, 119(4):e2115642118, 2022.

Elisabeth Dawson, Gonçalo R Abecasis, Suzannah Bumpstead, Yuan Chen, Sarah Hunt, David M Beare, Jagjit Pabial, Thomas Dibling, Emma Tinsley, Susan Kirby, et al. A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, 418(6897):544–548, 2002.

X. Didelot, G. Meric, D. Falush, and A. E. Darling. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics*, 13:256, Jun 2012.

B. Eldon, M. Birkner, J. Blath, and F. Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, Mar 2015.

Hans Ellegren. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.*, 29(1):51–63, 2014. ISSN 01695347. doi: 10.1016/j.tree.2013.09.008. URL `http://dx.doi.org/10.1016/j.tree.2013.09.008`.

Hans Ellegren, Nick GC Smith, and Matthew T Webster. Mutation rate variation in the mammalian genome. *Current Opinion in Genetics & Development*, 13(6):562–568, 2003. ISSN 0959-437X. doi: https://doi.org/10.1016/j.gde.2003.10.008. URL `https://www.sciencedirect.com/science/article/pii/S0959437X03001461`.

Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C. Sousa, and Matthieu Foll. Robust demographic inference from genomic and snp data. *PLOS Genetics*, 9(10):1–17, 10 2013. doi: 10.1371/journal.pgen.1003905. URL `https://doi.org/10.1371/journal.pgen.1003905`.

Adam Eyre-Walker and Peter D Keightley. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.*, 8(8):61061–8, 2007. ISSN 1471-0056. doi: 10.1038/nrg2146.

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

R Chris Gaynor, Gregor Gorjanc, and John M Hickey. AlphaSimR: an R package for breeding program simulations. *G3 Genes—Genomes—Genetics*, 11(2), 12 2020. ISSN 2160-1836. doi: 10.1093/g3journal/jkaa017. URL `https://doi.org/10.1093/g3journal/jkaa017`. jkaa017.

Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLOS Genetics*, 5(10):1–11, 10 2009. doi: 10.1371/journal.pgen.1000695. URL `https://doi.org/10.1371/journal.pgen.1000695`.

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

Chad Harland, Carole Charlier, Latifa Karim, Nadine Cambisano, Manon Deckers, Myriam Mni, Erik Mullaart, Wouter Coppieters, and Michel Georges. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. *bioRxiv*, 2017. doi: 10.1101/079863. URL `https://www.biorxiv.org/content/early/2017/06/29/079863`.

Kelley Harris and Rasmus Nielsen. The genetic cost of neanderthal introgression. *Genetics*, 203(2):881–891, 06 2016. ISSN 1943-2631. doi: 10.1534/genetics.116.186890. URL `https://doi.org/10.1534/genetics.116.186890`.

M Hartfield, N Aagaard Poulsen, B Guldbrandtsen, and T Bataillon. Using singleton densities to detect recent selection in bos taurus. *Evolution Letters*, 2022. doi: 10.1002/evl3.263. URL `https://doi.org/10.1002/evl3.263`.

Michael P Heaton, Timothy P L Smith, Derek M Bickhart, Brian L Vander Ley, Larry A Kuehn, Jonas Oppenheimer, Wade R Shafer, Fred T Schuetze, Brad Stroud, Jennifer C McClure, Jennifer P Barfield, Harvey D Blackburn, Theodore S Kalbfleisch, Kimberly M Davenport, Kristen L Kuhn, Richard E Green, Beth Shapiro, and Benjamin D Rosen. A Reference Genome Assembly of Simmental Cattle, Bos taurus taurus. *Journal of Heredity*, 112(2):184–191, 01 2021. ISSN 0022-1503. doi: 10.1093/jhered/esab002. URL `https://doi.org/10.1093/jhered/esab002`.

Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Bilal El Houdaigui, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Tiago Grego, Cristina Guijarro-Clarke, Leanne Haggerty, Anmol Hemrom, Thibaut Hourlier, Osagie G Izuogu, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, José Carlos Marugán, Thomas Maurel, Aoife C McMahon, Shamika Mohanan, Benjamin Moore, Matthieu Muffato, Denye N Oheh, Dimitrios Paraschas, Anne Parker, Andrew Parton, Irina Prosovetskaia, Manoj P Sakthivel, Ahamed I Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Emily Steed, Michal Szpak, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Brandon Walts, Andrea Winterbottom, Marc Chakiachvili, Ameya Chaubal, Nishadi De Silva, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth R IIsley, Nick Langridge, Jane E Loveland, Fergal J Martin, Jonathan M Mudge, Joanella Morales, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Fiona Cunningham, Andrew D Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa942. URL `https://doi.org/10.1093/nar/gkaa942`.

PingHsun Hsieh, Krishna R Veeramah, Joseph Lachance, Sarah A Tishkoff, Jeffrey D Wall, Michael F Hammer, and Ryan N Gutenkunst. 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, 2016.

Christian D. Huber, Arun Durvasula, Angela M. Hancock, and Kirk E. Lohmueller. Gene expression drives the evolution of dominance. *Nat. Commun.*, 9(1):2750, 2018. ISSN 20411723. doi: 10.1038/s41467-018-05281-7. URL `http://dx.doi.org/10.1038/s41467-018-05281-7`.

T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

P. D. Keightley, U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. L. Blaxter. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res*, 19 (7):1195–1201, Jul 2009.

Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

Katharine L. Korunes and Mohamed A. F. Noor. Gene conversion and linkage: effects on genome evolution and speciation. *Molecular Ecology*, 26(1):351–364, 2017. doi: https://doi.org/10.1111/mec.13736.

Christopher C. Kyriazis, Jacqueline A. Robinson, and Kirk E. Lohmueller. Using computational simulations to quantify genetic load and predict extinction risk. *bioRxiv*, 2022. doi: 10.1101/2022.08.12.503792. URL `https://www.biorxiv.org/content/early/2022/08/15/2022.08.12.503792`.

Greger Larson and Joachim Burger. A population genetics view of animal domestication. *Trends in Genetics*, 29(4):197–205, 2013.

Harris A. Lewin, Stephen Richards, Erez Lieberman Aiden, Miguel L. Allende, John M. Archibald, Miklós Bálint, Katharine B. Barker, Bridget Baumgartner, Katherine Belov, Giorgio Bertorelle, Mark L. Blaxter, Jing Cai, Nicolette D. Caperello, Keith Carlson, Juan Carlos Castilla-Rubio, Shu-Miaw Chaw, Lei Chen, Anna K. Childers, Jonathan A. Coddington, Dalia A. Conde, Montserrat Corominas, Keith A. Crandall, Andrew J. Crawford, Federica DiPalma, Richard Durbin, ThankGod E. Ebenezer, Scott V. Edwards, Olivier Fedrigo, Paul Flicek, Giulio Formenti, Richard A. Gibbs, M. Thomas P. Gilbert, Melissa M. Goldstein, Jennifer Marshall Graves, Henry T. Greely, Igor V. Grigoriev, Kevin J. Hackett, Neil Hall, David Haussler, Kristofer M. Helgen, Carolyn J. Hogg, Sachiko Isobe, Kjetill Sigurd Jakobsen, Axel Janke, Erich D. Jarvis, Warren E. Johnson, Steven J. M. Jones, Elinor K. Karlsson, Paul J. Kersey, Jin-Hyoung Kim, W. John Kress, Shigehiro Kuraku, Mara K. N. Lawniczak, James H. Leebens-Mack, Xueyan Li, Kerstin Lindblad-Toh, Xin Liu, Jose V. Lopez, Tomas Marques-Bonet, Sophie Mazard, Jonna A. K. Mazet, Camila J. Mazzoni, Eugene W. Myers, Rachel J. O'Neill, Sadye Paez, Hyun Park, Gene E. Robinson, Cristina Roquet, Oliver A. Ryder, Jamal S. M. Sabir, H. Bradley Shaffer, Timothy M. Shank, Jacob S. Sherkow, Pamela S. Soltis, Boping Tang, Leho Tedersoo, Marcela Uliano-Silva, Kun Wang, Xiaofeng Wei, Regina Wetzer, Julia L. Wilson, Xun Xu, Huanming Yang, Anne D. Yoder, and Guojie Zhang. The earth biogenome project 2020: Starting the clock. *Proceedings of the National Academy of Sciences*, 119(4):e2115635118, 2022. doi: 10.1073/pnas.2115635118. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2115635118`.

X. Liu and Y. X. Fu. Corrigendum: Exploring population size changes using SNP frequency spectra. *Nat Genet*, 47(9):1099, Sep 2015.

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

Xin Ma, Joanna L. Kelley, Kirsten Eilertson, Shaila Musharoff, Jeremiah D. Degenhardt, André L. Martins, Tomas Vinar, Carolin Kosiol, Adam Siepel, Ryan N. Gutenkunst, and Carlos D. Bustamante. Population genomic analysis reveals a rich speciation and demographic history of orang-utans (Pongo pygmaeus and Pongo abelii). *PLoS One*, 8(10):e77175, oct 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0077175. URL `http://dx.plos.org/10.1371/journal.pone.0077175`.

I M MacLeod, D M Larkin, H A Lewin, B J Hayes, and M E Goddard. Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. *Molecular Biology and Evolution*, 30(9):2209–2223, 07 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst125. URL `https://doi.org/10.1093/molbev/mst125`.

I M MacLeod, B J Hayes, and M E Goddard. The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. *Genetics*, 198(4):1671–1684, 09 2014. ISSN 1943-2631. doi: 10.1534/genetics.114.168344. URL `https://doi.org/10.1534/genetics.114.168344`.

B O Makanjuola, F Miglior, E A Abdalla, C Maltecca, F S Schenkel, and C F Baes. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of holstein and jersey cattle populations. *Journal of Dairy Science*, 2020. doi: 10.3168/jds.2019-18013. URL `https://doi.org/10.3168/jds.2019-18013`.

G. McVicker, D. Gordon, C. Davis, and P. Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet*, 5(5):e1000471, May 2009.

T H E Meuwissen, B J Hayes, and M E Goddard. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157(4):1819–1829, 04 2001. ISSN 1943-2631. doi: 10.1093/genetics/157.4.1819. URL `https://doi.org/10.1093/genetics/157.4.1819`.

A. Miles, N. J. Harding, G. Botta, C. S. Clarkson, T. Antao, K. Kozak, D. R. Schrider, A. D. Kern, S. Redmond, I. Sharakhov, R. D. Pearson, C. Bergey, M. C. Fontaine, M. J. Donnelly, M. K. N. Lawniczak, D. P. Kwiatkowski, M. J. Donnelly, D. Ayala, N. J. Besansky, A. Burt, B. Caputo, A. Della Torre, M. C. Fontaine, H. C. J. Godfray, M. W. Hahn, A. D. Kern, D. P. Kwiatkowski, M. K. N. Lawniczak, J. Midega, D. E. Neafsey, S. O'Loughlin, J. Pinto, M. M. Riehle, I. Sharakhov, K. D. Vernick, D. Weetman, C. S. Wilding, B. J. White, A. D. Troco, J. Pinto, A. Diabaté, S. O'Loughlin, A. Burt, C. Costantini, K. R. Rohatgi, N. J. Besansky, N. Elissa, J. Pinto, B. Coulibaly, M. M. Riehle, K. D. Vernick, J. Pinto, J. Dinis, J. Midega, C. Mbogo, P. Bejon, C. S. Wilding, D. Weetman, H. D. Mawejje, M. J. Donnelly, D. Weetman, C. S. Wilding, M. J. Donnelly, J. Stalker, K. Rockett, E. Drury, D. Mead, A. Jeffreys, C. Hubbart, K. Rowlands, A. T. Isaacs, D. Jyothi, C. Malangone, P. Vauterin, B. Jeffery, I. Wright, L. Hart, K. Kluczy?ski, V. Cornelius, B. MacInnis, C. Henrichs, R. Giacomantonio, D. P. Kwiatkowski, V. Cornelius, B. MacInnis, C. Henrichs, R. Giacomantonio, and D. P. Kwiatkowski. Genetic diversity of the African malaria vector Anopheles gambiae. *Nature*, 552(7683):96–100, 12 2017.

Valeria Montano. Coalescent inferences in conservation genetics: Should the exception become the rule? *Biol. Lett.*, 12(6), 2016. ISSN 1744957X. doi: 10.1098/rsbl.2016.0211.

Michael W. Nachman. Variation in recombination rate across the genome: Evidence and implications. *Curr. Opin. Genet. Dev.*, 12(6):657–663, 2002. ISSN 0959437X. doi: 10.1016/S0959-437X(02)00358-1.

Miguel Navascués and Brent C Emerson. Elevated substitution rate estimates from ancient dna: model violation and bias of bayesian methods. *Molecular Ecology*, 18(21):4390–4397, 2009.

Dominic Nelson, Jerome Kelleher, Aaron P. Ragsdale, Claudia Moreau, Gil McVean, and Simon Gravel. Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLOS Genetics*, 16 (5):1–12, 05 2020. doi: 10.1371/journal.pgen.1008619. URL `https://doi.org/10.1371/journal.pgen.1008619`.

J Obšteter, J Jenko, and G Gorjanc. Genomic selection for any dairy breeding program via optimized investment in phenotyping and genotyping. *Frontiers in Genetics*, 12, 2021. doi: 10.3389/fgene.2021.637017. URL `https://www.frontiersin.org/article/10.3389/fgene.2021.637017`.

March Pombi, Aram D. Stump, Allesandra Della Torre, and Nora J. Besansky. Variation in recombination rate across the x chromosome of anopheles gambiae. *The American Journal of Tropical Medicine and Hygiene*, 75(5):901–903, 2006. doi: https://doi.org/10.4269/ajtmh.2006.75.901. URL `https://www.ajtmh.org/view/journals/tpmd/75/5/article-p901.xml`.

Aaron P. Ragsdale, Dominic Nelson, Simon Gravel, and Jerome Kelleher. Lessons learned from bugs in models of human history. *The American Journal of Human Genetics*, 107(4):583–588, 2020. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2020.08.017. URL `https://www.sciencedirect.com/science/article/pii/S000292972030286X`.

Arang Rhie, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, William Chow, Arkarachai Fungtammasan, Juwan Kim, Chul Lee, Byung June Ko, Mark Chaisson, Gregory L. Gedman, Lindsey J. Cantin, Francoise Thibaud-Nissen, Leanne Haggerty, Iliana Bista, Michelle Smith, Bettina Haase, Jacquelyn Mountcastle, Sylke Winkler, Sadye Paez, Jason Howard, Sonja C. Vernes, Tanya M. Lama, Frank Grutzner, Wesley C. Warren, Christopher N. Balakrishnan, Dave Burt, Julia M. George, Matthew T. Biegler, David Iorns, Andrew Digby, Daryl Eason, Bruce Robertson, Taylor Edwards, Mark Wilkinson, George Turner, Axel Meyer, Andreas F. Kautt, Paolo Franchini, H. William Detrich III, Hannes Svardal, Maximilian Wagner, Gavin J. P. Naylor, Martin Pippel, Milan Malinsky, Mark Mooney, Maria Simbirsky, Brett T. Hannigan, Trevor Pesout, Marlys Houck, Ann Misuraca, Sarah B. Kingan, Richard Hall, Zev Kronenberg, Ivan Sović, Christopher Dunn, Zemin Ning, Alex Hastie, Joyce Lee, Siddarth Selvaraj, Richard E. Green, Nicholas H. Putnam, Ivo Gut, Jay Ghurye, Erik Garrison, Ying Sims, Joanna Collins, Sarah Pelan, James Torrance, Alan Tracey, Jonathan Wood, Robel E. Dagnew, Dengfeng Guan, Sarah E. London, David F. Clayton, Claudio V. Mello, Samantha R. Friedrich, Peter V. Lovell, Ekaterina Osipova, Farooq O. Al-Ajli, Simona Secomandi, Heebal Kim, Constantina Theofanopoulou, Michael Hiller, Yang Zhou, Robert S. Harris, Kateryna D. Makova, Paul Medvedev, Jinna

Hoffman, Patrick Masterson, Karen Clark, Fergal Martin, Kevin Howe, Brian P. Flicek, Paul Walenz, Woori Kwak, Hiram Clawson, Mark Diekhans, Luis Nassar, Benedict Paten, Robert H. S. Kraus, Andrew J. Crawford, M. Thomas P. Gilbert, Guojie Zhang, Byrappa Venkatesh, Robert W. Murphy, Klaus-Peter Koepfli, Beth Shapiro, Warren E. Johnson, Federica Di Palma, Tomas Marques-Bonet, Emma C. Teeling, Tandy Warnow, Jennifer Marshall Graves, Oliver A. Ryder, David Haussler, Stephen J. O'Brien, Jonas Korlach, Harris A. Lewin, Kerstin Howe, Eugene W. Myers, Richard Durbin, Adam M. Phillippy, and Erich D. Jarvis. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856):737–746, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03451-0.

Benjamin D Rosen, Derek M Bickhart, Robert D Schnabel, Sergey Koren, Christine G Elsik, Elizabeth Tseng, Troy N Rowan, Wai Y Low, Aleksey Zimin, Christine Couldrey, Richard Hall, Wenli Li, Arang Rhie, Jay Ghurye, Stephanie D McKay, Françoise Thibaud-Nissen, Jinna Hoffman, Brenda M Murdoch, Warren M Snelling, Tara G McDaneld, John A Hammond, John C Schwartz, Wilson Nandolo, Darren E Hagen, Christian Dreischer, Sebastian J Schultheiss, Steven G Schroeder, Adam M Phillippy, John B Cole, Curtis P Van Tassell, George Liu, Timothy P L Smith, and Juan F Medrano. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, 9(3), 03 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa021. URL `https://doi.org/10.1093/gigascience/giaa021`. giaa021.

Stephan Schiffels and Ke Wang. *MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent*, pages 147–166. Springer US, New York, NY, 2020. ISBN 978-1-0716-0199-0. doi: 10.1007/978-1-0716-0199-0_7. URL `https://doi.org/10.1007/978-1-0716-0199-0_7`.

Daniel R Schrider. Background selection does not mimic the patterns of genetic diversity produced by selective sweeps. *Genetics*, 216(2):499–519, 2020.

Daniel R. Schrider and Andrew D. Kern. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends Genet.*, 34(4):301–312, 2018. ISSN 13624555. doi: 10.1016/j.tig.2017.12.005. URL `http://dx.doi.org/10.1016/j.tig.2017.12.005`.

M. V. Sharakhova, M. P. Hammond, N. F. Lobo, J. Krzywinski, M. F. Unger, M. E. Hillenmeyer, R. V. Bruggner, E. Birney, and F. H. Collins. Update of the Anopheles gambiae PEST genome assembly. *Genome Biol*, 8(1):R5, 2007.

Fran Supek and Ben Lehner. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair*, 81:102647, 2019. ISSN 1568-7864. doi: https://doi.org/10.1016/j.dnarep.2019.102647. URL `https://www.sciencedirect.com/science/article/pii/S1568786419302009`. Cutting-edge Perspectives in Genomic Maintenance VI.

A Talenti, J Powell, J D Hemmink, E A J Cook, D Wragg, S Jayaraman, E Paxton, C Ezeasor, E T Obishakin, E R Agusi, A Tijjani, K Marshall, A Fisch, B R Ferreira, A Qasim, U Chaudhry, P Wiener, P Toye, L J Morrison, T Connelley, and J G D Prendergast. A cattle graph genome incorporating global breed diversity. *Nature Communications*, 2022. doi: 10.1038/s41467-022-28605-0. URL `https://doi.org/10.1038/s41467-022-28605-0`.

João C. Teixeira and Christian D. Huber. The inflated significance of neutral genetic diversity in conservation genetics. *Proc. Natl. Acad. Sci. U. S. A.*, 118(10):1–10, 2021. ISSN 10916490. doi: 10.1073/pnas.2015096118.

J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, and J. M. Akey. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, Jul 2012.

Kosuke M. Teshima, Graham Coop, and Molly Przeworski. How reliable are empirical genomic scans for selective sweeps? *Genome Res.*, 16(6):702–712, 2006. ISSN 10889051. doi: 10.1101/gr.5105206.

P M VanRaden. Symposium review: How to implement genomic selection. *Journal of Dairy Science*, 103(6):5291–5301, 2020. ISSN 0022-0302. doi: https://doi.org/10.3168/jds.2019-17684. URL `https://www.sciencedirect.com/science/article/pii/S002203022030309X`.

S. Wielgoss, J. E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming, C. Medigue, R. E. Lenski, and D. Schneider. Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli. *G3 (Bethesda)*, 1(3):183–186, Aug 2011.

Liangbiao Zheng, Mark Q Benedict, Anton J Cornel, Frank H Collins, and Fotis C Kafatos. An integrated genetic map of the african human malaria vector mosquito, anopheles gambiae. *Genetics*, 143(2):941–952, 1996.