Accessible, realistic genome simulation with selection using `stdpopsim`

Graham Gower[1,*], Nathaniel S. Pope[2,*], Murillo F. Rodrigues[2,3,*], Silas Tittes[2,*], Linh N. Tran[4,*], Ornob Alam[5], Maria Izabel A. Cavassim[6], Peter D. Fields[7], Benjamin C. Haller[8], Xin Huang[9,10], Ben Jeffrey[11], Kevin Korfmann[2], Christopher C. Kyriazis[6], Jiseon Min[2], Inés Rebollo[12], Clara T. Rehmann[2], Scott T. Small[2], Chris C. R. Smith[2], Georgia Tsambos[13], Yan Wong[11], Yu Zhang[14], Christian D. Huber[15], Gregor Gorjanc[14], Aaron P. Ragsdale[16], Ilan Gronau[17], Ryan N. Gutenkunst[4], Jerome Kelleher[11], Kirk E. Lohmueller[6], Daniel R. Schrider[18], Peter L. Ralph[2,19], and Andrew D. Kern[2,♡]

[*]These authors contributed equally to the paper.
[♡]To whom correspondence should be addressed.
[1]Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Denmark
[2]Institute of Ecology and Evolution, University of Oregon, Eugene, OR, 97402, USA
[3]Division of Genetics, Oregon National Primate Center, Oregon Health and Science University, Beaverton, OR, 97006, USA
[4]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ, 85721, USA
[5]Center for Genomics & Systems Biology, New York University, New York, NY, 10003, USA
[6]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA, USA
[7]The Jackson Laboratory, Bar Harbor, ME, 04609, USA
[8]Department of Computational Biology, Cornell University, Ithaca, NY, USA
[9]Department of Evolutionary Anthropology, University of Vienna, Vienna, Austria
[10]Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Vienna, Austria
[11]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, UK
[12]Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN, 55108, USA
[13]Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA
[14]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh EH25 9RG, UK
[15]Department of Biology, Pennsylvania State University, University Park, PA, 16802, USA
[16]Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI, USA
[17]Efi Arazi School of Computer Science, Reichman University, Herzliya 4610101, Israel
[18]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA
[19]Department of Data Science, University of Oregon, Eugene, OR, 97402, USA

March 23, 2025

Abstract

Selection is a fundamental evolutionary force that shapes patterns of genetic variation across species. However, simulations incorporating realistic selection along heterogeneous genomes in complex demographic histories are challenging, limiting our ability to benchmark statistical methods aimed at detecting selection and to explore theoretical predictions. `stdpopsim` is a community-maintained simulation library that already provides an extensive catalog of species-specific population genetic models. Here we present a major extension to the `stdpopsim` framework that enables simulation of various modes of selection, including background selection, selective sweeps, and arbitrary distributions of fitness effects (DFE) acting on annotated

subsets of the genome (for instance, exons). This extension maintains stdpopsim's core principles of reproducibility and accessibility while adding support for species-specific genomic annotations and published DFE estimates. We demonstrate the utility of this framework by benchmarking methods for demographic inference, DFE estimation, and selective sweep detection across several species and scenarios. Our results demonstrate the robustness of demographic inference methods to selection on linked sites, reveal the sensitivity of DFE-inference methods to model assumptions, and show how genomic features, like recombination rate and functional sequence density, influence power to detect selective sweeps. This extension to stdpopsim provides a powerful new resource for the population genetics community to explore the interplay between selection and other evolutionary forces in a reproducible, low-barrier framework.

Introduction

Selection is a fundamental force in evolution, shaping the genetic diversity of populations and driving the adaptation of species to their environments. The effects of selection on genetic variation are complex, and can be difficult to disentangle from other evolutionary processes such as mutation, recombination, and genetic drift (e.g., Gillespie, 1991). For instance, changes in population size can lead to fluctuations in genetic diversity across a recombining chromosome that can mimic the effects of selection (Simonsen, Churchill, and Aquadro, 1995), leading to spurious inferences about the strength and targets of genetic adaptation (Simonsen, Churchill, and Aquadro, 1995; Akey et al., 2004; Nielsen et al., 2005). In turn, selection can confound our ability to infer demographic history from allele frequencies (Ewing and Jensen, 2016; Schrider, Shanku, and Kern, 2016) and estimates of inverse coalescent rate (Schrider, Shanku, and Kern, 2016; Johri, Riall, et al., 2021; Cousins et al., 2024). Thus, it is imperative to jointly account for the effects of selection and demography when inferring evolutionary history from genetic data (e.g., Sheehan and Song, 2016; Johri, Charlesworth, and Jensen, 2020), yet the number of tools for doing so remain limited.

To meet the growing need for the interpretation, analysis, and exploration of realistic and complex evolutionary models, the field of population genetics has increasingly turned to simulation. Simulation in population genetics has a long history including both backward-in-time coalescent simulations (Kingman, 1982; Hudson, 1983; Hudson et al., 1990) and forward-in-time simulations of complex demography and selection (e.g., Gillespie, 1984; Hernandez, 2008; Thornton, 2014; Haller and Messer, 2019; Gaynor, Gorjanc, and Hickey, 2021). The development of simulation tools has been driven by several needs: understanding the effects of complex evolutionary processes on genetic variation (e.g., Peischl et al., 2013; Thornton, 2019; Galloway, Cresko, and Ralph, 2020; Zhang, Kim, Lohmueller, et al., 2020), obtaining a null model for hypothesis testing (e.g., Hudson, Boos, and Kaplan, 1992; Hudson, Bailey, et al., 1994; Sabeti et al., 2002), exploring the power and limitations of statistical methods (e.g., Przeworski, 2002; MacLeod, Hayes, and Goddard, 2014), and increasingly, providing a basis for machine learning and other simulation-based inference methods (e.g., Beaumont, Zhang, and Balding, 2002; Lin et al., 2011; Kern and Schrider, 2018; Mughal and DeGiorgio, 2019; Sanchez et al., 2021; Wang et al., 2021; Zhang, Kim, Singh, et al., 2023; Smith and Kern, 2023). However, joint simulation of complex demography and selection on realistic genomes is challenging, requiring many choices including the strength and rate of selection, and a parameterized demographic model, as well as implementation of genetic maps and other genomic features related to the strength of selection. This is a daunting task, and can be a barrier to the adoption of simulation-based methods in population genetics. Furthermore, different simulation and inference tools often use different structures and parameter scalings for their models, making it difficult to directly compare across these tools and the studies that use them.

In light of the complexities associated with estimating and simulating population genetic models, it is no surprise that a lingering challenge with simulation in population genetics has been reproducibility and the ability to share and compare results among researchers (e.g., Ragsdale, Nelson, et al., 2020). This challenge has been addressed in part by the development of community resources for sharing and distributing simulation software via the stdpopsim project (Adrion et al., 2020). stdpopsim provides a standardized interface for accessing a wide range of population genetic models, and has been widely adopted by the community (e.g., Speidel et al., 2021; Wang et al., 2021; Yang et al., 2022; DeGiorgio and Szpiech, 2022; Browning, Waples, and Browning, 2023; Schweiger and Durbin, 2023; Temple, Waples, and Browning, 2024; Haag, Jordan, and Stamatakis, 2025). Yet, previous releases of stdpopsim did not include models of selection,

which is a major limitation for more empirically motivated applications of population genetic simulation. In particular, modeling selection through simulation is critical for understanding processes such as adaptation (e.g. Thornton, 2019; Hartfield and Glémin, 2024), the effects of selective sweeps (e.g. Braverman et al., 1995; Fay and Wu, 2000; Przeworski, 2002; Przeworski, Coop, and Wall, 2005; Schrider, Mendes, et al., 2015), and the impact of background selection on genetic diversity (e.g. Charlesworth, Morgan, and Charlesworth, 1993; Charlesworth, Charlesworth, and Morgan, 1995; Williamson and Orive, 2002; Ewing and Jensen, 2016; Torres et al., 2020). Ideally, one would like to have a simple, unified framework for simulating both neutral and non-neutral evolutionary processes, and to be able to compare the results of these simulations to empirical data in a manner that is both accessible to a wide range of researchers and highly reproducible. Furthermore, the framework should include the complex realities of genomes, including heterogeneous recombination rates, variation in the size and density of functional elements, and non-equilibrium demographic histories. Moreover, one should be able to model all of these features simultaneously using estimates and annotations obtained from a species of interest (e.g. Schrider, 2020; Rodrigues, Kern, and Ralph, 2024). `stdpopsim` provides a framework for such models, thus lowering the barrier to production of complex, multifaceted simulations, and allowing a wider range of researchers to make use of high-quality simulations in developing or testing their work. A central goal of the `stdpopsim` project is open, collaborative, and inclusive work; we view capacity-building across the community as equal in importance to software development.

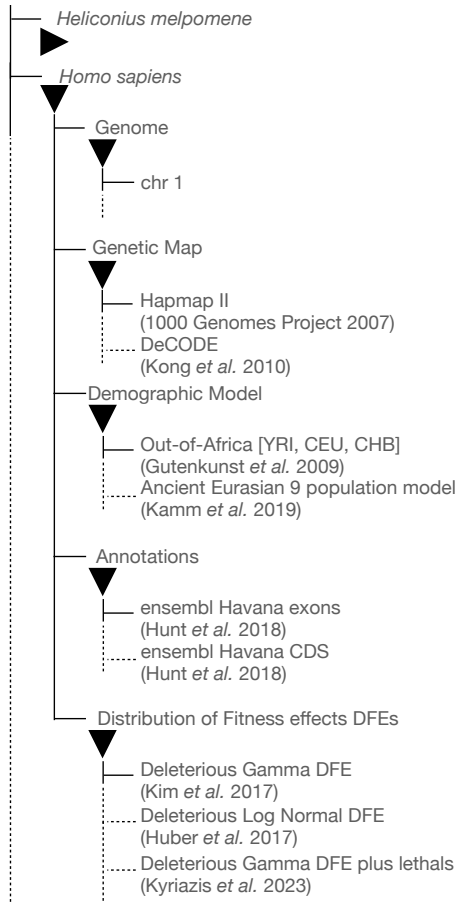
Here, we provide an overview of a major new addition to `stdpopsim`—the inclusion of models of selection. We begin by describing the models of selection that we have implemented, as well as the parameter choices that are available to the user. This includes models of background selection, selective sweeps, and models of the distribution of fitness effects (DFE). Further, we describe how these can be applied not only genome-wide but to genomic regions with a particular annotation, specific to a given species available through the `stdpopsim` application programming interface (API), and combined with a parameterized model of demography and recombination to provide a realistic simulation of genetic variation in a population. We then provide a series of examples of how these models can be used to benchmark and compare the performance of different methods for using genetic data to infer demographic history, the distribution of fitness effects, and the targets of recent positive selection.

Implementing Selection in `stdpopsim`

`stdpopsim` previously provided a wide range of models of neutral demographic history from roughly two dozen species (Lauterbur et al., 2023). This was accomplished through the use of a standardized interface and data structure that allows for the curated addition of new species and models, and can use either `msprime` (Baumdicker et al., 2021) or `SLiM` (Haller and Messer, 2019) as the backend engine for efficient population genetic simulation. To extend `stdpopsim`'s functionality to include models of selection, we introduce two new classes to the `stdpopsim` API: the `DFE` class and the `Annotation` class. The `DFE` class provides an interface for specifying a distribution of the fitness effects (DFE) of new mutations. Each `DFE` object specifies a list of mutation types and their proportions, where each mutation type is associated with a dominance coefficient, (currently limited to values between 0 and 1, inclusive), the distribution of selection coefficients (e.g., fixed-value, normally-distributed, gamma-distributed), and the parameters of this distribution. Each `Annotation` object represents a collection of functional genomic elements (e.g., exons, introns, intergenic regions, etc.) whose coordinates are specified using a GFF3 annotation file. More concretely, consider implementing a model of selection consistent with the DFE inferred by Kim et al. (2017) for exons in the human genome. The associated `DFE` object is defined with two mutation types, neutral and deleterious, corresponding to synonymous and non-synonymous sites, with their proportions set to their expected prevalence in exons. The neutral mutation type is associated with a fixed selection coefficient (0), and the deleterious mutation type is associated with a dominance coefficient of 0.5 and a gamma-distributed selection coefficient with mean and shape as inferred by Kim et al. (2017). `DFE` objects can also model a relationship between the selection and dominance coefficients (Huber, Durvasula, et al., 2018) by defining ranges of selection coefficients and associating each range with a different dominance coefficient (see below).

As with other aspects of `stdpopsim`, we intend the selection models to be as biologically accurate as possible. Thus, the `stdpopsim` catalog provides `Annotation` objects based on species' publicly available functional genomic elements and `DFE` objects based on published DFE estimates (Figure 1A). By using

A. Catalog



B. Python API

example_with_selection.py

```python
import stdpopsim

species = stdpopsim.get_species("HomSap")
contig = species.get_contig(
    "chr22",
    genetic_map="HapMap_GRCh37"
)
model = species.get_demographic_model(
    "OutOfAfrica_3G09"
)
samples = {"YRI": 5, "CEU": 5, "CHB": 5}
# Adding DFE to exons
dfe = species.get_dfe("Gamma_K17")
exons = species.get_annotations("ensembl_havana_104_exons")
exon_intervals = exons.get_chromosome_annotations("chr22")
contig.add_dfe(intervals=exon_intervals, DFE=dfe)
engine = stdpopsim.get_engine("slim")
# Performing the simulation
ts = engine.simulate(
    model,
    contig,
    samples,
    seed=236,
    slim_scaling_factor=1,
    slim_burn_in=10,
)
```

C. Command Line Interface

```
$ stdpopsim -e slim --slim-scaling-factor 20 HomSap \
    -c chr22 \
    --dfe Gamma_K17 \
    --dfe-annotation ensembl_havana_104_CDS \
    -o example.ts -d OutOfAfrica_3G09 YRI:5 CEU:5 CHB:5
```

Figure 1: **(A)** A schematic of the `stdpopsim` catalog and the different components that can be used in simulations. **(B)** The Python API and **(C)** the command line interface. The user specifies a species, a portion of the genome to simulate, and optionally a genetic map, a model of demography, and a model of selection that itself is composed of a distribution of fitness effects (DFE) and a set of functional annotations.

annotation objects, one can, for example, simulate populations where selected mutations (whose selection coefficients are drawn from a specified DFE) occur within annotation regions (e.g. exons), while all mutations occurring outside of these regions do not affect fitness. There are relatively few species with published DFE estimates in the literature. `stdpopsim` currently has implemented DFEs for five species in the catalog: *Arabidopsis thaliana* (Huber, Durvasula, et al., 2018), *Drosophila melanogaster* (Ragsdale, Coffman, et al., 2016; Huber, Kim, et al., 2017; Zhen et al., 2021), humans (Huber, Kim, et al., 2017; Kim, Huber, and Lohmueller, 2017; Kyriazis, Robinson, and Lohmueller, 2023; Castellano et al., 2019; Castellano et al., 2019; Rodrigues, Kern, and Ralph, 2024), *Mus musculus* (Booker et al., 2021), and the vaquita porpoise *Phocoena sinus* (Robinson et al., 2022). While the DFEs are from a limited number of species, these DFEs can be applied to other species in the catalog if and when it is suitable. For instance, one could simulate a butterfly species with a gamma-distributed DFE originally estimated from humans, or a human population with a DFE estimated from *Drosophila*. Indeed, the DFE labeled `HomSap/Mixed_K23` is recommended by Kyriazis, Robinson, and Lohmueller (2023) for "general" use across mammals (lacking more specific information). Furthermore, the user can specify a custom DFE and provide their own annotations of functional elements

to simulate selection in a species for which we do not yet have a published DFE included in the catalog.

From a user's perspective, a model of selection is specified by pairing a DFE object with a collection of genomic segments defined using an Annotation object. Multiple such pairings can be defined in one simulation to provide a rich model of how selection may vary along a chromosome. For example, coding sequences may be associated with one DFE and non-coding sequences in exons with another DFE. Simulations with selection are then implemented by specifying a species, a portion of the genome to simulate, a genetic map if available, a model of demography, and pairings of DFEs with genomic segments using either the Python API or the command-line interface (Figure 1B-C). Simulations with selection are generated using the SLiM simulation engine, which is perhaps the most flexible simulation engine for modeling selection available to date (msprime's selection models are relatively limited and thus not currently used by stdpopsim).

While the DFE class facilitates the simulation of non-neutral variation across a sequence, it does not provide a means of specifying the origin and fate of particular selected alleles. Thus, we have also implemented an API in stdpopsim (stdpopsim.selective_sweep) that enables selective sweeps to be simulated with a high degree of control. This function augments a demographic model by introducing an advantageous mutation over a given time interval, at a given position along the sequence; and conditions the simulation such that the mutation is above a particular frequency at the beginning and end of the sweep period. As this functionality is implemented on top of the existing stdpopsim API, it may be combined with other models of selection and demography to simulate the combined effects of multiple, disparate evolutionary processes on genetic variation in a population.

At this release (version 0.3.0) the stdpopsim catalog comprises 27 species (i.e. genome representations), including vertebrates, invertebrates, plants, and bacteria, for which we have included 35 demographic models, 38 genetic maps, 12 DFEs, and 8 annotation objects. Additions to the catalog are ongoing—we point the reader to our previously published report detailing this effort (Lauterbur et al., 2023). We welcome contributions from the community and encourage those interested in contributing to visit https://popsim-consortium.github.io/stdpopsim-docs or contact one of the PopSim Consortium members for guidance.

Inference of $N_e(t)$ in the context of selection

One of the most common applications of population genetic inference is to estimate the effective population size over time, $N_e(t)$, from genetic data. This can be done using a variety of methods, including the sequential Markovian coalescent (Li and Durbin, 2011; Schiffels and Wang, 2020; Terhorst, Kamm, and Song, 2017), through use of the site frequency spectrum (SFS) (Liu and Fu, 2020), as well as through linkage-disequilibrium information (Santiago et al., 2020). Since all of these methods assume neutrality, they are typically applied to genomic segments that are not expected to be directly affected by selection (e.g., by masking exons and conserved elements). However, selection acting on linked sites—even those that are masked out of the analysis—can bias estimates of $N_e(t)$ away from census population sizes by increasing the rate of coalescence in the genome (e.g. Schrider, Shanku, and Kern, 2016). Using stdpopsim, we can readily examine how different methods for inferring $N_e(t)$ are influenced by selection on linked sites, which in the case of negative selection is commonly referred to as *background selection* (Charlesworth, Morgan, and Charlesworth, 1993; Hudson and Kaplan, 1995).

To do this we simulated human genomes with and without selection. In both scenarios, we ran three replicate simulations under the out of Africa (OOA) demographic model of Ragsdale and Gravel (2019) using a genetic map from the HapMap Project (International HapMap Consortium et al., 2007) (see **Methods**). Simulations with selection were implemented by modeling purifying selection on nonsynonymous mutations using a DFE inferred by Kim, Huber, and Lohmueller (2017). A basic summary of nucleotide diversity from these simulations shows that the presence of selection decreases polymorphism only slightly at the genome-wide level, but more substantially within exonic regions (Figure S1). Following the simulation, inference of $N_e(t)$ was conducted from each of the six simulated datasets using four methods: stairwayplot (Liu and Fu, 2020), msmc2 (Schiffels and Wang, 2020), SMC++ (Terhorst, Kamm, and Song, 2017) and GONE (Santiago et al., 2020).

Although the estimates produced by the four methods differ from one another, for each method the estimates produced from data with selection (Figure 2; orange) and without selection (Figure 2; blue) are

fairly similar. Thus, this particular parameterization of background selection and demography, in combination with selection only at exonic sites, does not appear to notably bias the estimates produced by these methods. This makes sense in light of our observation that genetic diversity is mostly affected within exonic regions and presumably sites very closely linked (Figure S1), thus masking of exons should be sufficient to remove the major effects of selection. As noted in previous studies that compared demographic inference methods on neutral simulations (Adrion et al., 2020), we see that each method has its strengths and weaknesses. Methods based on the sequential Markovian coalescent (`SMC++` and `msmc2`) appear to produce the most accurate trajectories of $N_e(t)$ overall in this setting, although they tend to under-estimate the population bottleneck associated with the population split. On the other hand, `stairwayplot`, which utilizes the site frequency spectrum, produces more noisy estimates, but it does not appear to under-estimate the population bottleneck. The inference of `GONE`, which utilizes linkage-disequilibrium, is targeted toward demographic changes in more recent time frames (~ 200 generations), explaining its noisy estimates for all but the most recent time periods.

To expand these observations to additional species and to highlight the ease of comparison enabled by `stdpopsim`, we performed a similar benchmark using the vaquita porpoise *Phocoena sinus*. We simulated 100 genomes under a two-epoch demographic model inferred for vaquita by Robinson et al. (2022), and a constant recombination rate across the genome (see **Methods**). In simulations with selection, we applied the DFE inferred by Robinson et al. (2022) to exons annotated from the vaquita genome assembly. This DFE implements a relationship between selection and dominance in which very deleterious mutations (with selection coefficient $s < -0.1$) are fully recessive (dominance coefficient $h = 0$) and weakly deleterious mutations ($s \geq -0.001$) are nearly additive ($h = 0.4$; see **Methods**). As in the human simulations, we see that the presence of selection decreases polymorphism in genomes simulated with selection, but that the effect is only readily apparent within exonic regions (Figure S2). For these simulations, the SFS-based method (`stairwayplot`) produces more accurate inference than the two methods based on the sequential Markovian coalescent (`SMC++` and `msmc2`; see Figure S3). Regardless of these differences, we see that background selection again does not appear to considerably influence the inference of any of the four methods for this species, at

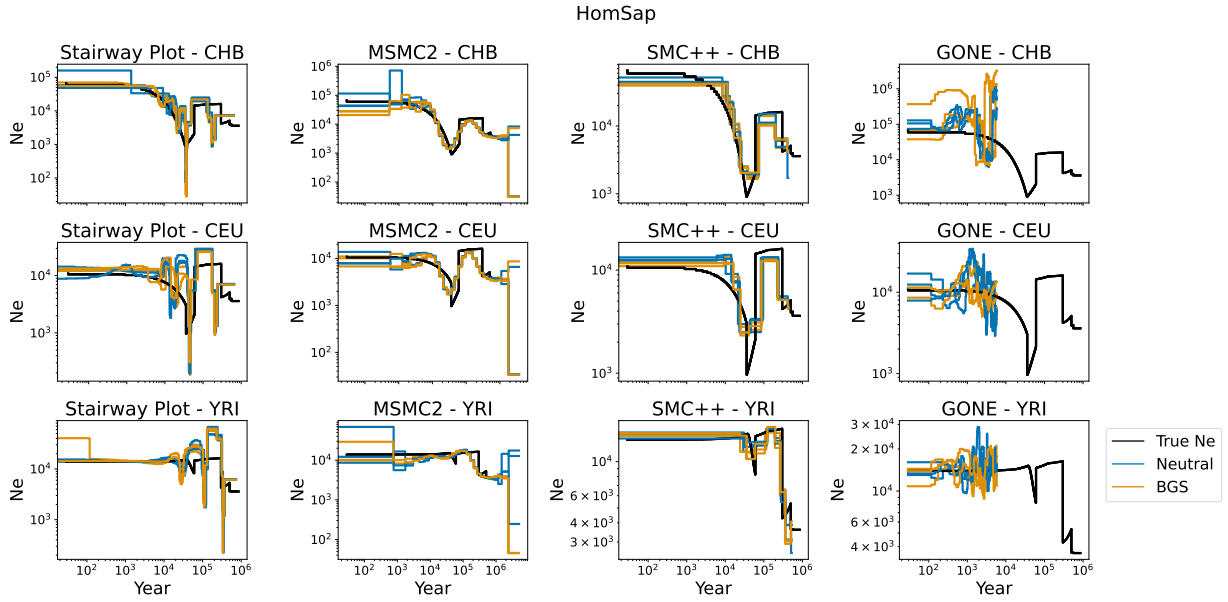


Figure 2: Inference of $N_e(t)$ from human genomes simulated under an out-of-Africa demographic model with and without purifying selection on exons. Rows correspond to the three extant populations in the simulation: CHB, CEU, and YRI. Columns correspond to the four inference methods: `stairwayplot`, `msmc2`, `SMC++` and `GONE`. Each plot depicts the inferred $N_e(t)$ on the three datasets simulated without selection (blue) and the three datasets simulated with the background effects of purifying selection on exons (orange), alongside the true values of N_e used in simulation (black).

least for the parameterization used here, even though vaquita has a very different evolutionary history than humans.

Estimation of the Distribution of Fitness effects

Another common application of population genetic inference is to estimate the distribution of fitness effects (DFE) of new mutations from genetic data. The new framework of modeling selection in stdpopsim makes it an ideal tool for easily benchmarking and comparing the performance of different methods for inferring the DFE. Here, we considered three DFE-inference methods: GRAPES (Galtier, 2016), polyDFE (Tataru and Bataillon, 2020), and dadi-cli (Huang et al., 2023).

A challenge when comparing different inferred and simulated distributions of the selection coefficient (s) is differing conventions in the literature regarding its definition. The simulator SLiM and the inference tool GRAPES define the selection coefficient s such that a homozygote has fitness $1+s$. On the other hand, dadi-cli and polyDFE define the selection coefficient such that a homozygote has fitness $1 + 2s$. Moreover, inferred distributions of selection coefficients are typically scaled by an inferred ancestral effective population size N_e, and different methods assume different scaling. dadi-cli infers the distribution of $2N_e s$ while polyDFE and GRAPES infer the distribution of $4N_e s$. To allow direct comparison, here we have normalized all inference results to the convention used by SLiM (also used by stdpopsim; see **Methods**).

We start by examining DFE inference from human genomes simulated under an out-of-Africa demographic model with selection acting on exons (see **Methods**). In this setting, the DFE is inferred separately for each of the three simulated populations using each of the three DFE-inference methods. In each case we note that the DFE is estimated from segregating sites only, without the use of substitution data, as stdpopsim's models currently do not contain outgroup populations or species that can be used to estimate divergence.

Overall, inferred DFEs varied across populations and inference methods (Figure 3, Table S1). Note that while the estimation accuracy for the DFE parameters themselves is informative (Figure 3A-B), examining the difference between the actual and inferred distributions of s can be more helpful for showing how well a method has recovered the true DFE (Figure 3C). The distributions of s inferred by polyDFE and dadi-cli had higher mean values (in absolute value) but smaller shape parameters, while GRAPES inferred distributions with lower mean values and shape parameters. As a result, polyDFE and dadi-cli over-estimate the fraction of mutations under strong selection ($s < -0.1$), while GRAPES considerably over-estimates the fraction of mutations under weak selection ($s \geq -0.0001$).

Population history appears to influence different methods in slightly different ways. For instance, dadi-cli and GRAPES produce more accurate estimates from the YRI genomes (relative to the other two populations), whereas polyDFE shows very similar accuracy across all populations. To factor out the influence of demographic changes, we also analyzed genomes generated using similar simulations but with no population size changes (Figure S4, Table S1). In this simpler setting, polyDFE overlapped the true mean selection coefficient and shape parameter, whereas the other methods do not show as close a correspondence. GRAPES exhibits similar biases as observed in the simulations with more complex demography, but with milder under-estimation of the shape parameter, whereas dadi-cli shows an inverse pattern with overestimated shape parameter and slightly underestimated mean. While providing possible insights into the effects of the demographic history on accuracy of different DFE-inference methods, substantiating these insights will require more than three replicate simulations per scenario.

Next, we benchmarked these three DFE-inference methods on simulations of the vaquita porpoise genome. Recall that these simulations apply a gamma-distributed selection coefficient to mutations in exons with a relationship between the selection and dominance coefficients (see **Methods**). In this setting, all methods perform uniformly worse than in the human genome simulations, with consistent underestimation of the mean selection coefficient and overestimation of the shape parameter (Figure 4A-B, Table S1). Similar patterns are observed when re-running the simulations with a simpler demographic model without population size changes (Figure S5, Table S1). We suspect that this is due to the fact that all DFE-inference methods used here assume an additive model ($h = \frac{1}{2}$), while mutations in the simulated genomes were more recessive ($h < \frac{1}{2}$), and their dominance decreased with the strength of selection. To examine the influence of this model misspecification, we compared the inferred distributions of s to the simulated distribution of $2hs$, which would have been equal to s under an additive model (Figure 4C). Reassuringly, all three methods

infer distributions that appear to fit the main mode of the simulated distribution of $2hs$. This is likely because deleterious alleles are typically at low frequency and thus heterozygous, where their selective effect is hs rather than s.

Detection of Selective Sweeps

We next highlight how `stdpopsim` can be used to benchmark and compare the performance of different methods for detecting selective sweeps from genetic data. Because simulation of selective sweeps is very time consuming and evaluation of statistical power requires many replicate simulations, we focused on human chromosome 1 and generated replicate simulations on 5cM segments of the chromosome in 93 different locations. Selective sweeps were simulated by introducing a beneficial mutation in the center of the 5cM segment with a moderately strong selection coefficient ($s = 0.03$; $2N_e s \sim 600$; see **Methods**). Each segment was simulated 1000 times, while keeping only replicates in which the selected allele reached a frequency of 95% or greater. These sweeps were modeled in the context of the three population out-of-Africa model from Gutenkunst et al. (2009) and background selection from deleterious mutations in exons. Sweep-detection methods typically base their inference on checking whether a test statistic exceeds a certain threshold, which is determined according to some empirical null distribution. In our simulations, we considered two different

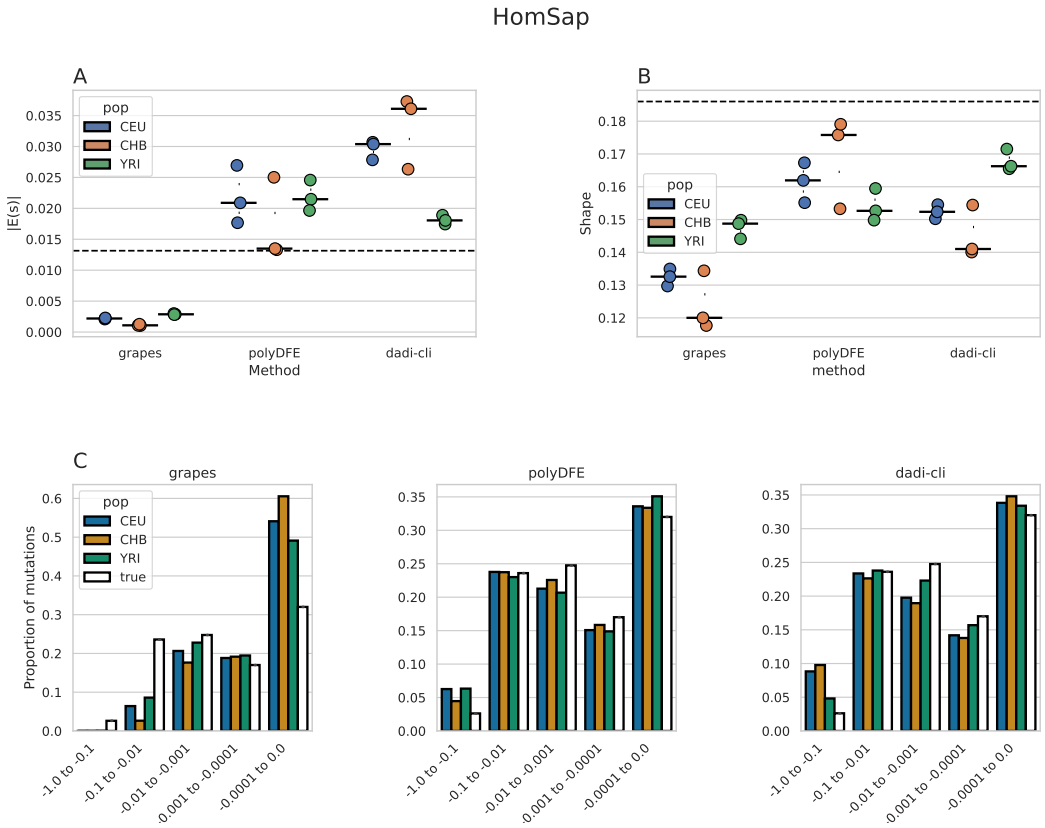


Figure 3: Inferred versus simulated DFEs from simulated human genomes. Simulations were performed using a human out-of-Africa demographic model with a gamma-distributed DFE acting on exons (see **Methods**). DFE is inferred separately for each of the three extant populations (CEU, CHB, and YRI) by three different methods: `GRAPES`, `polyDFE`, and `dadi-cli`. (A-B) Mean absolute value of selection coefficient ($|E(s)|$) and shape parameter are shown for each DFE inferred from all three simulated datasets, with median values marked by horizontal bars and simulated values represented by dashed horizontal lines. (C) Binned distribution of s implied by the average DFE inferred from the three simulated datasets (averaging the inferred gamma parameters); white bars represent the distribution used in the simulation.

null distributions: one based on neutral simulations, and another based on simulations with background selection from deleterious mutations in exons (see **Methods**).

We evaluated the statistical power to detect selective sweeps for three methods: `sweepfinder2` (DeGiorgio, Huber, et al., 2016), `diploshic` (Kern and Schrider, 2018), and reduced diversity (π). For each method, we determined a threshold for its test statistic using each of the two null models separately – neutral and with background selection (BGS) – and controlling for a 5% false positive rate. We then applied each method to data simulated with sweeps for two samples (YRI and CEU) in each of the simulated locations along human chromosome 1 (Figure 5). First and most strikingly, the power to detect sweeps varies considerably along the chromosome, and this variation is much greater than the variation in power between CEU and YRI or between null models. This is likely due to the effects of recombination rate heterogeneity (Figure 5). Next,

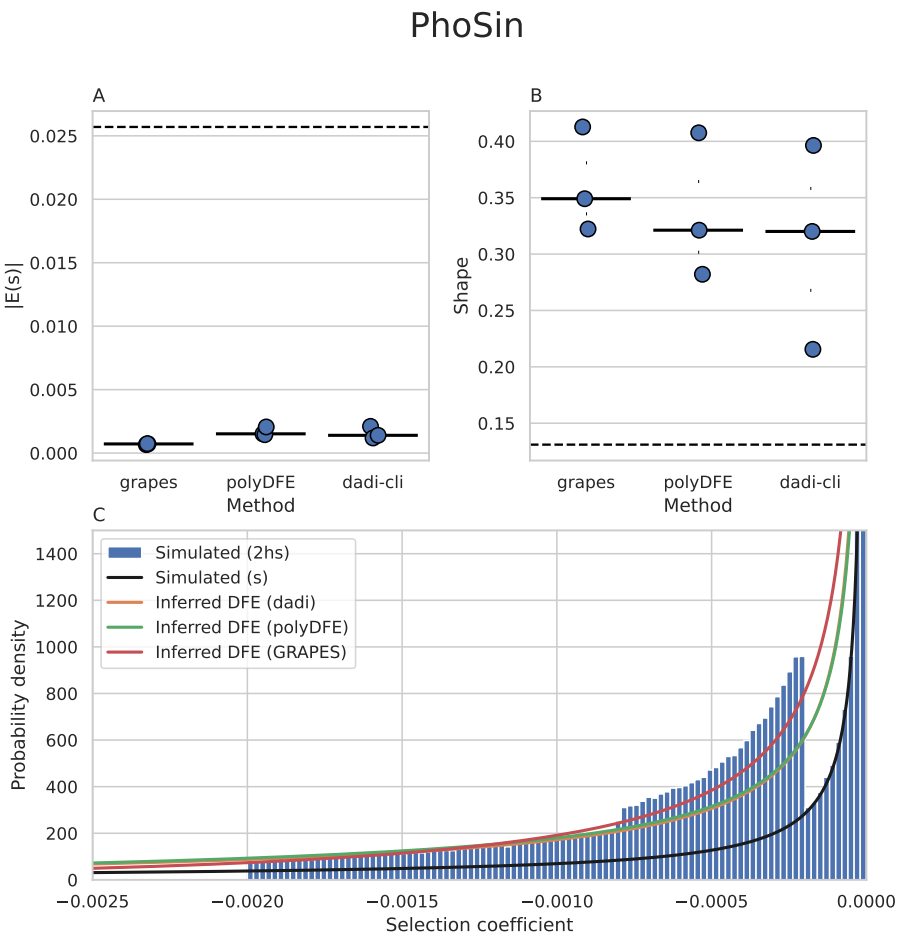


Figure 4: Inferred versus simulated DFEs from simulated vaquita genomes. Simulations were performed using a two-epoch model of vaquita porpoise demography with a gamma-distributed DFE acting on nonsynonymous mutations with a relationship between the selection coefficient (s) and dominance coefficient (h) (see **Methods**). The DFE is inferred by analyzing all simulated genomes jointly by one of three different methods: `GRAPES`, `polyDFE`, and `dadi-cli`. (A-B) Mean absolute value of selection coefficient ($|E(s)|$) and shape parameter are shown for each DFE inferred from all three simulated datasets, with median values marked by horizontal bars and simulated values represented by dashed horizontal lines. (C) Binned distribution of $2hs$ implied by the simulated DFE compared with the distribution of s implied by the average DFE inferred for each method from the three simulated datasets (averaging the inferred gamma parameters). The distribution of $2hs$ is multimodal because of the simulated relationship between h and s (see text).

we notice that power to detect sweeps is slightly lower when the test statistic threshold is determined using simulated data with background selection (BGS null model). This is likely because the strength of selection acting on exons is relatively weak, and the sweep detection methods we have used are relatively robust to

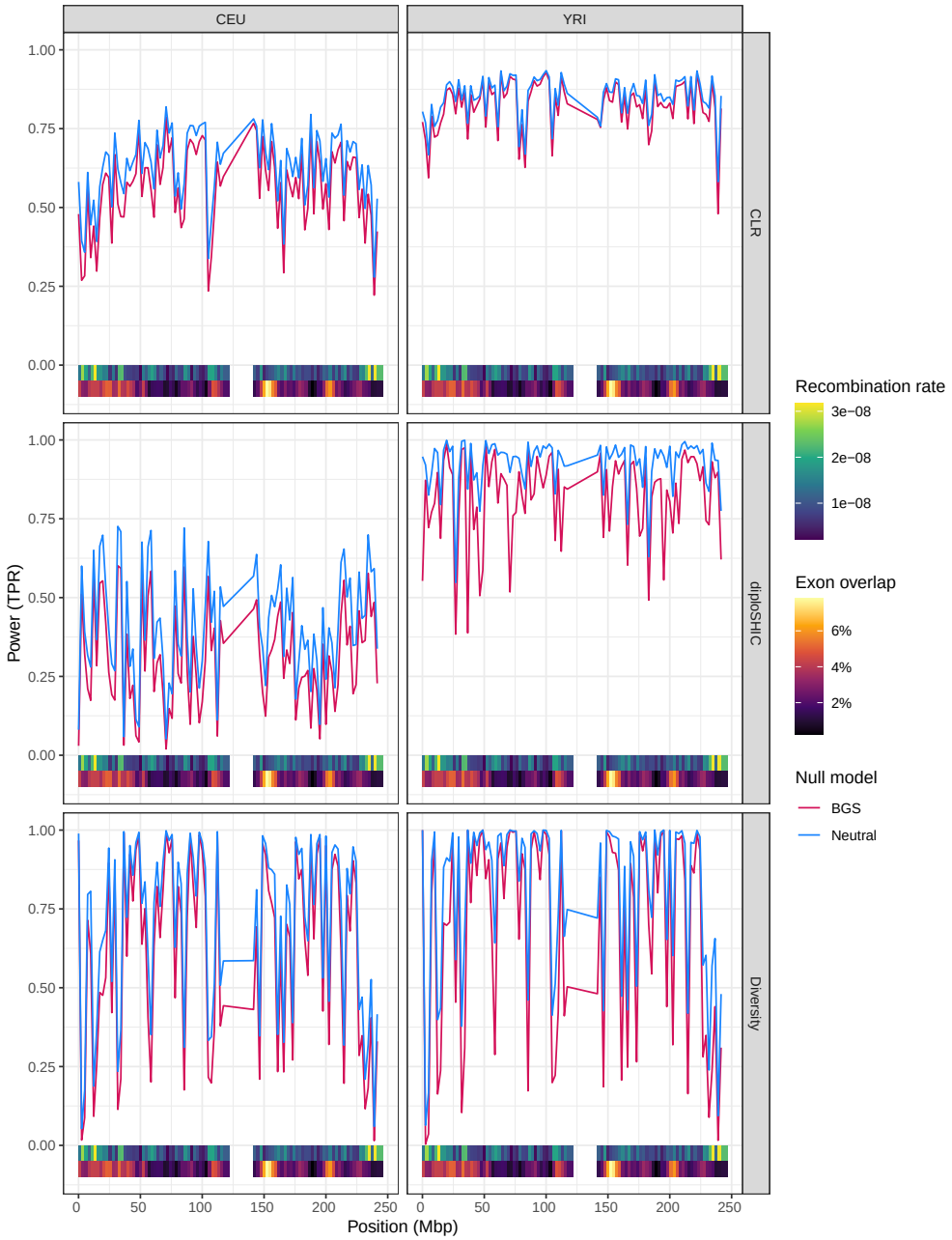


Figure 5: Power to detect selective sweeps along human chromosome 1. Genomic segments were simulated with sweeps under a three population out-of-Africa model and with background selection from deleterious mutations in exons. Three methods for detecting sweeps were applied to simulated data: `sweepfinder2` (top row – labeled CLR), `diploshic` (middle row), and reduced diversity (π) (bottom row). Power (true positive rate) is shown for these methods for the CEU and YRI samples (left and right respectively). The thresholds of the test statistics were set to control for 5% false positive rate under a neutral null model (blue) and a null model with background selection from deleterious mutations in exons (red). Also shown are heatmaps of local recombination rate and exon density along the chromosome (bottom).

the effects of background selection. Finally, the power to detect sweeps is generally higher in the YRI sample than in the CEU sample, which is consistent with the fact that YRI has a larger ancestral effective population size and thus both stronger selection and more variation to detect a sweep (e.g., Simonsen, Churchill, and Aquadro, 1995). Conversely, the CEU population experienced the out-of-Africa bottleneck that may mute the signal of a sweep.

We compared the overall performance of the three sweep-detection methods using receiver operating characteristic (ROC) curves (Figure S6). Performance varies for each of the sampled populations. In the CEU sample, `sweepfinder2` and `diploshic` perform similarly with a high area under the curve (AUC) score, (AUC = 0.906 and 0.903 respectively), but both underperform the simple method that uses reduced diversity (π) as an indicator (AUC = 0.932). In the YRI sample, `diploshic` outperforms the other two methods, with an AUC of 0.979, followed by using outliers of π with an AUC of 0.957, and then `sweepfinder2` with an AUC of 0.939. We note that we trained `diploshic` on simulations under a constant size population, which more closely matches the demographic history of YRI, than that of CEU, which is characterized by a strong recent bottleneck followed by population expansion. Interestingly, the simple method that uses reduced diversity (π) as an indicator for a selective sweep performs very well overall (especially on the CEU sample). However, its accuracy is much more variable along the chromosome compared to the other two methods (see Figure 5), and for quite a few locations along the chromosome it has power below 50% when applied to both sample sets. `sweepfinder2` appears to have the least variation in power along the chromosome or across populations.

To further dissect the reasons for high variance along the chromosome in power to detect sweeps, we examined the relationship between power and local recombination rate (Figure 6). We see that power to detect sweeps is a decreasing function of local recombination rate for `sweepfinder2` and reduced diversity, which is consistent with sweeps having a larger genomic footprint in regions of low recombination. `diploshic` on the other hand shows a slight increase in power with recombination rates, presumably because it is designed to distinguish regions where a sweep happened from regions linked to a sweep, and if the recombination rate is too low (or selection too strong) then the genomic window examined by `diploshic` may not be large enough to exhibit enough of a recovery in patterns of diversity flanking the sweep. These results confirm that reduced genetic diversity performs well as an indicator for selective sweeps only in regions with low recombination rate. The other two methods, which were specifically designed for sweep detection, are considerably more robust to the effects of recombination rate than the simple indicator of reduced genetic diversity. These methods are particularly robust to recombination rate variation when applied to the YRI sample, especially when using a neutral null model to calibrate the test statistics. The other null model, which assumes background selection from exons, appears to consistently lead to reduced power for detecting sweeps regardless of the recombination rate. Interestingly, we do not see a clear correlation between power to detect sweeps and local exon density (Figure S7). This is somewhat surprising given that background selection is expected to somewhat interfere with the sweep signature. This might suggest that the effects of realistic background selection on genome variation may not be as similar to those of sweeps as previously thought (Schrider, 2020).

Discussion

In this paper, we have presented an important new addition to the `stdpopsim` library that enables the simulation of genetic variation in the presence of selection. We have demonstrated the utility of this new API by showing how it can be used to benchmark the performance of different methods for demographic history inference, the inference of the distribution of fitness effects of new mutations, and the detection of selective sweeps. This allows for a wide range of models of selection to be easily simulated and compared to empirical data, with reproducibility, computational efficiency, and rigor.

The exploration of performance among methods done here is not meant to be thorough benchmarking, but rather to highlight the utility of the `stdpopsim` resource. Having said that, there are a number of interesting results that merit discussion. The first pertains to estimation of demographic history ($N_e(t)$) in the context of linked purifying selection. Our simulations of human and vaquita genomes model deleterious mutations only at exonic positions of the genome. As shown in Figures S1 and S2, this combination only produces sizable decreases in polymorphism at exonic positions, which represent a small minority of sites

within the genome, and does not lead to large-scale perturbations in genome-wide diversity. Thus, it is not surprising that estimates of $N_e(t)$ from genome-wide data from which exons have been masked should be accurate (e.g., Figures 2 and S3). However, the question remains to what extent this would be true in empirical data where selection is not constrained to coding positions. For instance, we know that many genomes harbor large numbers of conserved non-coding elements (e.g., Siepel et al., 2005) with extremely strong selection acting on them (Katzman et al., 2007; McLean and Bejerano, 2008). We thus caution regarding extrapolation from our results. Future studies might build upon the annotations we used here to include selection at other portions of the genome as well as mixtures of DFEs across sites.

A second important finding is that inference of the DFE is sensitive to the dominance of deleterious mutations. All three of the DFE-inference methods applied here assume that deleterious mutations act in an

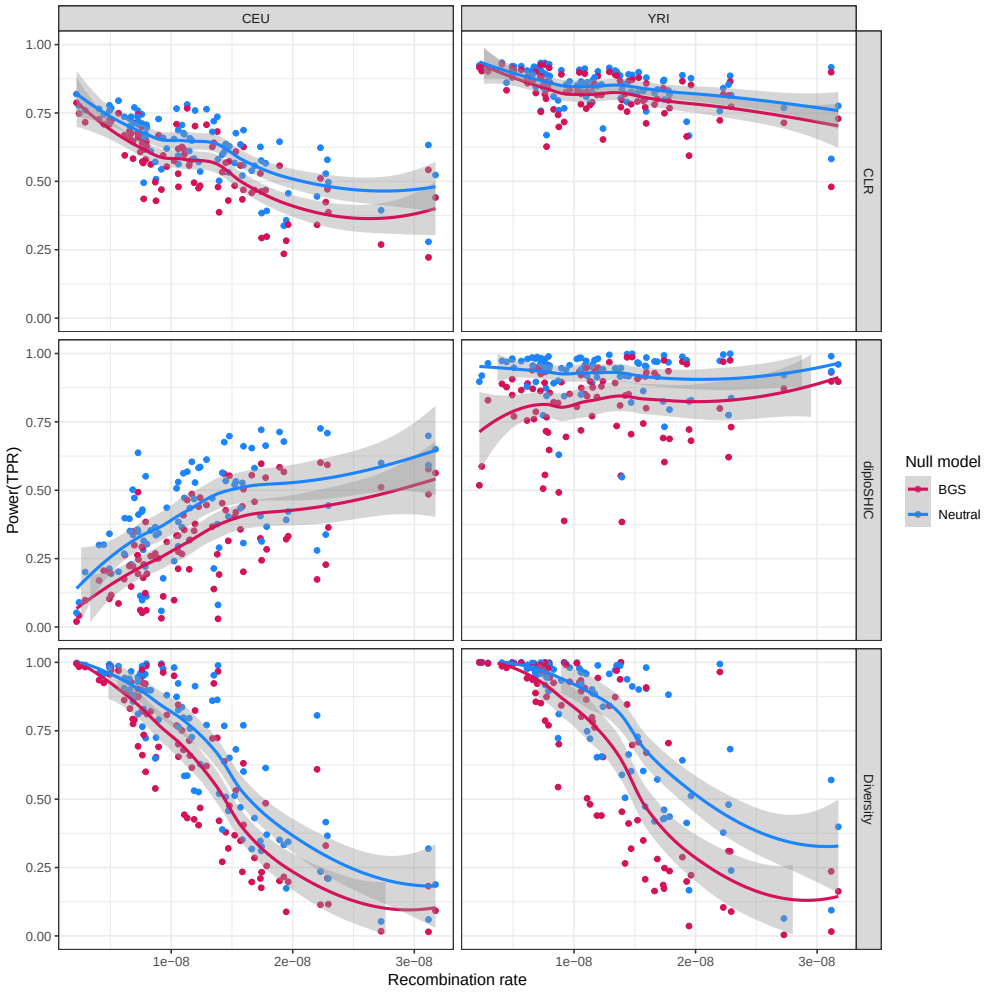


Figure 6: Power to detect selective sweeps as a function of local recombination rate. This figure shows the same power estimates shown in Figure 5, but with the genomic segments plotted against their average recombination rates instead of position along chromosome 1. Genomic segments were simulated with sweeps under a three population out-of-Africa model and with background selection from deleterious mutations in exons. Three methods for detecting sweeps were applied to simulated data: `sweepfinder2` (top row– labeled CLR), `diploshic` (middle row), and reduced diversity (π) (bottom row). Power (true positive rate) is shown for these methods for the CEU and YRI samples (left and right respectively). The thresholds of the test statistics were set to obtain a 5% false positive rate under a neutral null model (blue) and a null model with background selection from deleterious mutations in exons (red). Fitted lines represent loess smoothed regressions.

additive fashion ($h = 0.5$) by default. However, the simulated DFE for the vaquita genome included highly recessive mutations. Specifically, these simulations used a DFE where mutations with $s < -0.01$ are quite recessive, with $h <= 0.01$, and where mutations with $-0.01 < s < -0.001$ have $h = 0.1$. We found that the three DFE-inference methods underestimated $E[s]$ and overestimated the shape parameter (Figure 4A-B). This performance is not surprising since decreasing h for a given distribution of s will lead to a decrease in hs, or an apparent decrease in the amount of selection. Fitting a mis-specified DFE that assumes $h = 0.5$ will then lead to underestimates of $E[s]$. Indeed, the distribution of hs is much more accurately estimated (Figure 4C). Overall, these results are consistent with prior work showing that it is challenging to separately infer the DFE of s along with h (Veeramah et al., 2014; Kyriazis and Lohmueller, 2024; Balick et al., 2022) and that a large proportion ($> 5\%$) of recessive deleterious mutations can confound inferences of s (Wade et al., 2023). Given the plethora of evidence for recessive deleterious mutations (Mukai et al., 1972; Agrawal and Whitlock, 2011; Huber, Durvasula, et al., 2018; Di and Lohmueller, 2024), these results suggest the need for care in DFE inference. Kyriazis and Lohmueller 2024 inferred the DFE of s for nonsynonymous mutations in humans when considering different values of h. Intriguingly, additive models showed a satisfactory fit to the SFS of nonsynonymous variants in humans, and more complex models of dominance did not substantially improve the fit. Nevertheless, models including recessive deleterious mutations, like those simulated here for the vaquita genome, show a similar fit to the SFS and may be more biologically realistic than fully additive models, given the evidence for the presence of recessive deleterious mutations. More broadly, new computational approaches are needed to precisely quantify the DFE and the covariance of h and s.

Further, inference of the DFE was done here using exon annotations of the genome in question. A key distinction between our analyses and those done in empirical work is that we have complete knowledge of which sites are under selection in the simulation. Our inferences therefore represent a best case scenario. Empirical inference of the DFE is a much harder problem, since some mutations assumed to be free of selection might in reality be under selection, and the strength of selection can change over time or across environments.

Regarding sweep detection, the inclusion of selection models into `stdpopsim` allowed us to readily compare the effectiveness of different methods using data that reflects the demographic history, recombination rate variation, and genomic locations of negatively selected sites present in human populations. These results clearly demonstrate the impact of recombination rate variation, and underscore the importance of training methods on data with variable recombination rates (e.g., Schrider and Kern, 2017). In addition, our results illustrate a strength of likelihood methods like `sweepfinder2` that do not require a predefined window size. Window-based methods, such as `diploshic`, tend to lag in performance when the genomic region affected by the sweep is larger than the chosen window size, as can occur with very strong sweeps or low levels of recombination. Moreover, while all the methods examined here retain adequate power for sweep identification across scenarios, including the inclusion of background selection, our results underscore the importance of developing methods with adequate statistical power even in the face of unknown demographic histories. Indeed, larger differences in power were observed when applying a method to genomes simulated with different demographic histories (i.e., YRI versus CEU) than when calibrating a method using a null model with and without background selection.

Although we have shown how the new `stdpopsim` API can be used to benchmark the performance of different methods, the benchmarks shown here merely scratch the surface of what can be done with the new API. We envision that, moving forward, many researchers will use the new API to simulate data under a wide range of models of selection, demography, and genome architecture, and to develop new methods for the analysis of genetic data in the presence of selection. The additions described here further expand the utility of `stdpopsim` for both characterizing the patterns of diversity expected under a wide range of population genetic models, and for benchmarking methods for the analysis of genetic data on common ground in a way that was not possible before.

Methods

Annotations

Currently available annotation data in `stdpopsim` were downloaded from Ensembl. Coding sequence and exon annotations are available for *Arabidopsis thaliana* (Cheng et al., 2017), *Drosophila melanogaster*

(Hoskins et al., 2015), *Homo sapiens* (Hunt et al., 2018a), and *Phocoena sinus* (vaquita porpoise, Morin et al., 2021).

Quality control and software development

Following the protocol established previously for other aspects of stdpopsim (Adrion et al., 2020), new DFEs are subject to a rigorous quality control (QC) process, by which a different individual re-implements the same DFE, and the two implementations are programmatically tested for equality. This QC process has proven its effectiveness on numerous occasions in detecting implementation bugs and resolving modeling ambiguities. For example, this process detected an error in coding the log-normal distribution of the *Drosophila melanogaster* DFE inferred by Booker et al. (2021, stdpopsim label LognormalPlusPositive_R16). This faulty DFE model was used in a study conducted by one of our groups before it underwent the QC process, where we noticed that deleterious selection coefficients were several orders of magnitude larger than expected. Subsequently, the QC process was able to find and fix this bug. In addition to QC, all changes and additions to the code must pass a comprehensive set of unit tests: at the time of writing, the library includes over 11,000 lines of testing code, comprising 2,862 unit tests, including comparisons to QC implementations and cross-checks for agreement, including between SLiM and msprime implementations of the same models. For instance, for testing we make use of SLiM's scriptability to separately record as metadata the simulation parameters (current population sizes, DFEs, etc.), which are then programmatically compared to the models used to generate the SLiM script that ran the simulation. We make use of automatic test "code coverage" tools to help ensure that all portions of the code base are validated by these unit tests.

Whole genome simulations

In this study we performed simulations of the human and vaquita porpoise genomes using stdpopsim. We ran replicate simulations for each genome/parameter combination, in which chromosomes were simulated separately from one another. All simulations were executed using the SLiM engine with a burn-in of $10N$ generations (where N is the initial population size), and a scaling factor of 1 (no scaling). We note that aggressive scaling of forward-in-time simulations can lead to quite different outcomes when selection in involved Uricchio and Hernandez, 2014; Ferrari et al., 2024; Dabi and Schrider, 2025, and we caution users against using high degrees of scaling for drawing final inferences, although we recommend it during development to save on computational costs. All simulations were implemented using a snakemake workflow, where the code utilized is similar to that shown in Figure 1B. The complete simulation workflow is available at https://github.com/popsim-consortium/analysis2/blob/main/workflows/simulation.snake (see **Data availability** below).

Simulations of complete human genomes

Simulation of human genomes used a genetic map from the HapMap Project (International HapMap Consortium et al., 2007, stdpopsim label HapMapII_GRCh38) and the out of Africa demographic model with archaic admixture from Ragsdale and Gravel (2019, stdpopsim label OutOfAfricaArchaicAdmixture_5R19). Simulations with selection additionally applied a DFE model with gamma-distributed selection coefficients inferred by Kim, Huber, and Lohmueller (2017, stdpopsim label Gamma_K17) to exons annotated by the HAVANA group release 104 (Hunt et al., 2018b, stdpopsim label ensembl_havana_104_exons). Each simulation generated 100 diploid genomes from each of three extant human populations (labeled YRI, CEU, and CHB), with each genome consisting of 22 autosomes. Using these settings, we generated three replicate datasets without selection and three replicate datasets with selection. Three additional datasets, used in Figure S4, were simulated with the same model of selection and a simple demographic model with one population with a constant size of $10,000$. Each of these simulations generated 100 diploid genomes. The configuration file used in combination with our snakemake workflow for these simulations can be found at https://github.com/popsim-consortium/analysis2/blob/main/workflows/config/snakemake/production_config_HomSap.yml.

Simulations of complete vaquita porpoise genomes

Simulation of vaquita porpoise genomes used the two-epoch demographic model from Robinson et al. (2022, stdpopsim label Vaquita2Epoch_1R22). Since no genetic map is currently available for this species, we assumed a constant recombination rate of 10^{-8} across all chromosomes (Morin et al., 2021). Simulations with selection additionally applied a DFE model with gamma-distributed selection coefficients inferred by Robinson et al. (2022, stdpopsim label Gamma_R22) to exons annotated from the vaquita porpoise genome assembly Morin et al. (2021, stdpopsim label Phocoena_sinus.mPhoSin1.pri.110_exons). This DFE model implements a relationship between selection and dominance as follows: very deleterious mutations (with $s < -0.1$) are recessive ($h = 0$), mutations with $s \in [-0.1, -0.01)$ have a dominance coefficient of $h = 0.01$, mutations with $s \in [-0.01, -0.001)$ have a dominance coefficient of $h = 0.1$, and very mild mutations (with $s \in [-0.001, 0)$) are nearly additive ($h = 0.4$). This discretized relationship matches that used by Robinson et al. (2022), and is easily implemented in stdpopsim's DFE class. Each simulation generates 100 diploid genomes, with each genome consisting of 21 autosomes. Using these settings, we generated three replicate datasets without selection and three replicate datasets with selection. Three additional datasets, used in Figure S5, were simulated with the same model of selection and a simple demographic model with one population with a constant size of $3,500$. The configuration file used in combination with our snakemake workflow for these simulations can be found at https://github.com/popsim-consortium/analysis2/blob/main/workflows/config/snakemake/production_config_PhoSin.yml.

Inference of $N_e(t)$

Inference of $N_e(t)$ was conducted on simulated datasets after masking non-recombining portions of the genome (to avoid bias caused by extreme recombination rate variation) and exons (to avoid the direct effect of natural selection). Thus, inference from datasets simulated with selection was not influenced by direct selection acting on exons, but it was influenced by selection acting on linked sites (background selection). We assessed the accuracy of demographic inferences from four software packages: msmc2 (Schiffels and Wang, 2020), stairwayplot (Liu and Fu, 2020), GONE (Santiago et al., 2020), and SMC++ (Terhorst, Kamm, and Song, 2017). For msmc2 we used a random sample of $N = 6$ simulated genomes with 20 iterations of the EM algorithm. For GONE we set max_snps to 500,000, the number of generations to 2,000, and the number of bins to 400, all representing default settings. For stairwayplot we did one run of θ estimation for each replicate dataset, and used a dimFactor of 5000, both of which represent default settings. For SMC++ we used the unfolded SFS (--unfold), along with the true mutation rate under which the simulations were generated. The complete snakemake workflow for demographic inference can be found at https://github.com/popsim-consortium/analysis2/blob/main/workflows/n_t.snake.

DFE inference

We used three software packages to infer the distribution of fitness effects (DFE) from genetic data: dadi-cli (Huang et al., 2023; Kim, Huber, and Lohmueller, 2017), polyDFE (Tataru and Bataillon, 2020), and GRAPES (Galtier, 2016). Each software package was run using the same set of simulations. DFE inference using dadi-cli was conducted by first inferring a demographic model from neutral sites, using a two epoch demographic model with 100 repeated runs of the optimization procedure. dadi-cli then proceeds by using a cached strategy wherein we used 2000 points in the gamma distribution. Finally, the DFE was inferred using a ratio of non-synonymous to synonymous mutations of 2.31, again with 100 repeated runs of the optimization procedure. For polyDFE, we used model 'C' which uses a mixture of a gamma distribution and an exponential distribution, where the gamma component is used for negatively selected mutations. For optimization we used the default BFGS optimizer with the -e flag to automatically estimate the initial parameters values. For GRAPES, we inferred the DFE under the 'GammaZero' model, which assumes that selection coefficients are drawn from a single gamma distribution. We used the -no-div-param flag to disable the use of divergence-based parameter estimation. All other settings were left at their default values. All methods used the same exon annotations as generated in the simulations. The total number of coding sites in the human and vaquita genomes used were 61,644,801 and 31,039,513 respectively.

To facilitate direct comparison between inference methods, we scale each program's output to obtain a selection coefficient (s) consistent with the one used by SLiM. GRAPES outputs the mean and shape of the

inferred distribution of $4N_e s$. We thus scale the mean by a factor of $\frac{\mu}{\theta}$, where μ is the assumed mutation rate and $\theta = 4N_e\mu$ is the population-scaled mutation rate, as estimated by GRAPES. polyDFE is similarly scaled by $\frac{2\mu}{\theta}$, reflecting the fact that it defines the selection coefficient such that a homozygote has fitness $1 + 2s$. dadi-cli infers a gamma distribution for $N_e s$ and parameterizes this distribution using the shape (α) and rate (ν). Thus, the mean value of s is obtained by $\frac{4\mu}{\theta}\alpha\nu Lp$, where L is the total number of sites available for the analysis, and p is the proportion of sites under selection. Complete details for the DFE inference analysis can be found in the snakemake workflow at https://github.com/popsim-consortium/analysis2/blob/main/workflows/dfe.snake.

Simulations with selective sweeps

To test methods for detecting selective sweeps, we considered 100 evenly distributed 5cM segments of human chromosome 1. As recombination rate varies across the genome, the size of the region in base pairs simulated varied, however the recombination distance, which is critical for the linked effects of a sweep, was held constant. Seven segments in which the central 1Mb had an average recombination rate below 10^{-11} were excluded from the analysis. Each segment was simulated in each of the following three settings: (1) neutral, (2) background selection (BGS), and (3) hard sweep with BGS. In all three settings, we applied a genetic map from the HapMap Project (International HapMap Consortium et al., 2007, stdpopsim label HapMapII_GRCh38) and the three-population out of Africa (OOA) demographic model from Gutenkunst et al. (2009, stdpopsim label OutOfAfrica_3G09). In the two settings with BGS, we applied the DFE model with gamma-distributed selection coefficients inferred by Kim, Huber, and Lohmueller (2017, stdpopsim label Gamma_K17) to exons annotated by the HAVANA group release 104 (Hunt et al., 2018b, stdpopsim label ensembl_havana_104_exons). Hard sweeps were modeled by generating a single beneficial mutation with selection coefficient of $s = 0.03$ in the center of the 5cM genomic segment and discarding simulations in which the beneficial mutation did not reach (or exceed) a terminal frequency of 95%. For each of the 93 genomic segments of chromosome 1, we generated 1,000 replicate datasets under each of the three settings. We used the SLiM engine with a burn-in of $2N$ generations (where N is the initial population size), and a scaling factor of 2. Each simulation generated ten diploid genomic segments for each of the two sampled populations: YRI and CEU. The sequences simulated with a hard sweep and BGS were used for sweep detection, and sequences simulated under the other two settings were used to calibrate the thresholds of the test statistics used by the three sweep-detection methods (see below).

Selective sweeps detection

Each of the three sweep-detection methods is based on a test statistic. Reduced nucleotide diversity is based on nucleotide diversity (π) computed directly from the underlying tree sequences using tskit (Ralph, Thornton, and Kelleher, 2020) over ten equally sized and non-overlapping subwindows. sweepfinder2 (DeGiorgio, Huber, et al., 2016) is based on a composite likelihood ratio (CLR) computed at 21 uniformly distributed positions along the simulated region. diploshic (Kern and Schrider, 2018) uses supervised learning to train a sweep classifier. Its training phase used discoal (Kern and Schrider, 2016) to simulate 20 haploid sequences of length 1.21Mbp with a constant population size of $N_e = 10,000$, a mutation rate of 2×10^{-8}, and a recombination rate of 2×10^{-8}. For discoal simulations that included selective sweeps, selection coefficients were drawn from a uniform distribution between 0.001 and 0.10, and the time of the sweep was chosen from a uniform distribution between 0 and $0.01N_e$ generations in the past. For soft sweeps, the frequency at which selection was introduced was drawn from a uniform distribution between 0.0001 and 0.1. We note that since the training set of diploshic was generated under a simple model with constant mutation and recombination rates and constant effective population size, we expect it to perform suboptimally on the simulated segments of human chromosome 1, which were generated using more realistic simulations (see **Simulations with selective sweeps** above).

The three sweep-detection methods were associated with thresholds for their respective test statistics (π, sweepfinder2's CLR, or diploshic's probability), which were determined using an empirical distribution representing a null model without sweeps. We considered two null models: a neutral evolution model (neutral), and a model with background selection from deleterious mutations in exons (BGS). Each null model was represented by a different set of $93,000$ simulations generated according to the appropri-

ate model (see **Simulations with selective sweeps** above). Thus, for each test statistic we determined two thresholds corresponding to these two different null models (neutral and BGS) using the extreme 5% quantile in each empirical distribution. Then, power (or true positive rate) to detect a sweep by a given method was determined for positions on chromosome 1 as follows: each sweep-detection method was applied to the 1,000 simulated datasets generated with a hard sweep and BGS for that position, and the power was set to be the proportion of simulations (out of 1,000) in which the test statistic exceeded the threshold determined by each of the two null models (neutral and BGS). We note that using an empirical distribution to determine the threshold for calling a selective sweep with `diploshic` deviates from the original implementation of the method, however for consistency among methods we chose to take this approach. The complete `snakemake` workflow for sweep detection (including simulations) can be found at `https://github.com/popsim-consortium/analysis2/blob/main/workflows/sweep_simulate.snake`.

Data availability

The release of `stdpopsim` is available at `https://github.com/popsim-consortium/stdpopsim`. This paper coincides with version 0.3.0 of `stdpopsim`. The entire analysis pipeline associated with this paper is available as a `snakemake` workflow at `https://github.com/popsim-consortium/analysis2`.

Acknowledgments

We thank Tom Booker for discussions and encouragement to incorporate realistic BGS models into `stdpopsim`. We thank the numerous users of the `stdpopsim` package, as well as workshop participants who provided feedback and suggestions for improvements.

Funding

M. Rodgrigues, S. Tittes, N. Pope, and A. Kern were supported in part by NIH award R35GM148253. R. Gutenkunst and L. Tran were supported in part by NIH award R35GM149235. D. Schrider was supported in part by NIH award R35GM138286. K. Lohmueller was supported by NIH award R35GM119856. P. Ralph, B. Haller, J. Kelleher, and A. Kern were supported in part by NIH award R01HG012473. G. Gorjanc and Y. Zhang were supported in part by BBSRC (BBS/E/RL/230001A, BBS/E/RL/230001C, BB/T014067/1), Chinese Scolarship Council (CSC) and Genus. C Huber was supported by NIH award R35GM146886.

References

Adrion, Jeffrey R et al. (June 2020). "A community-maintained standard library of population genetic models". In: *eLife* 9. Ed. by Graham Coop et al., e54967. ISSN: 2050-084X. DOI: `10.7554/eLife.54967`. URL: `https://doi.org/10.7554/eLife.54967`.

Agrawal, Aneil F and Michael C Whitlock (2011). "Inferences about the distribution of dominance drawn from yeast gene knockout data". In: *Genetics* 187.2, pp. 553–566.

Akey, Joshua M et al. (2004). "Population history and natural selection shape patterns of genetic variation in 132 genes". In: *PLoS biology* 2.10, e286.

Balick, Daniel J et al. (2022). "Overcoming constraints on the detection of recessive selection in human genes from population frequency data". In: *The American Journal of Human Genetics* 109.1, pp. 33–49.

Baumdicker, Franz et al. (Dec. 2021). "Efficient ancestry and mutation simulation with msprime 1.0". In: *Genetics* 220.3. iyab229. ISSN: 1943-2631. DOI: `10.1093/genetics/iyab229`. eprint: `https://academic.oup.com/genetics/article-pdf/220/3/iyab229/43780247/iyab229.pdf`. URL: `https://doi.org/10.1093/genetics/iyab229`.

Beaumont, Mark A, Wenyang Zhang, and David J Balding (2002). "Approximate Bayesian computation in population genetics". In: *Genetics* 162.4, pp. 2025–2035.

Booker, Tom R. et al. (2021). "Selective sweeps influence diversity over large regions of the mouse genome". In: *bioRxiv*. DOI: 10.1101/2021.06.10.447924. eprint: https://www.biorxiv.org/content/early/2021/06/11/2021.06.10.447924.full.pdf. URL: https://www.biorxiv.org/content/early/2021/06/11/2021.06.10.447924.

Braverman, John M et al. (1995). "The hitchhiking effect on the site frequency spectrum of DNA polymorphisms." In: *Genetics* 140.2, pp. 783–796.

Browning, Sharon R, Ryan K Waples, and Brian L Browning (2023). "Fast, accurate local ancestry inference with FLARE". In: *The American Journal of Human Genetics* 110.2, pp. 326–335.

Castellano, David et al. (Nov. 2019). "Comparison of the Full Distribution of Fitness Effects of New Amino Acid Mutations Across Great Apes". In: *Genetics* 213.3, pp. 953–966. ISSN: 1943-2631. DOI: 10.1534/genetics.119.302494. eprint: https://academic.oup.com/genetics/article-pdf/213/3/953/49496002/genetics0953.pdf. URL: https://doi.org/10.1534/genetics.119.302494.

Charlesworth, Brian, MT Morgan, and Deborah Charlesworth (1993). "The effect of deleterious mutations on neutral molecular variation." In: *Genetics* 134.4, pp. 1289–1303.

Charlesworth, Deborah, Brian Charlesworth, and MT Morgan (1995). "The pattern of neutral molecular variation under the background selection model." In: *Genetics* 141.4, pp. 1619–1632.

Cheng, Chia-Yi et al. (2017). "Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome". In: *The Plant Journal* 89.4, pp. 789–804. DOI: https://doi.org/10.1111/tpj.13415. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.13415. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.13415.

Cousins, Trevor et al. (2024). "Accurate inference of population history in the presence of background selection". In: *bioRxiv*.

Dabi, Amjad and Daniel R Schrider (2025). "Population size rescaling significantly biases outcomes of forward-in-time population genetic simulations". In: *Genetics* 229.1, iyae180.

DeGiorgio, Michael, Christian D Huber, et al. (2016). "SweepFinder2: increased sensitivity, robustness and flexibility". In: *Bioinformatics* 32.12, pp. 1895–1897.

DeGiorgio, Michael and Zachary A Szpiech (2022). "A spatially aware likelihood test to detect sweeps from haplotype distributions". In: *PLoS genetics* 18.4, e1010134.

Di, Chenlu and Kirk E Lohmueller (2024). "Revisiting dominance in population genetics". In: *Genome Biology and Evolution* 16.8, evae147.

Ewing, Gregory B and Jeffrey D Jensen (2016). "The consequences of not accounting for background selection in demographic inference". In: *Molecular ecology* 25.1, pp. 135–141.

Fay, Justin C and Chung-I Wu (2000). "Hitchhiking under positive Darwinian selection". In: *Genetics* 155.3, pp. 1405–1413.

Ferrari, Tessa et al. (2024). "Towards simulation optimization: an examination of the impact of scaling on coalescent and forward simulations". In: *bioRxiv*, pp. 2024–04.

Galloway, Jared, William A Cresko, and Peter Ralph (2020). "A few stickleback suffice for the transport of alleles to new lakes". In: *G3: Genes, Genomes, Genetics* 10.2, pp. 505–514.

Galtier, Nicolas (2016). "Adaptive protein evolution in animals and the effective population size hypothesis". In: *PLoS genetics* 12.1, e1005774.

Gaynor, R C, G Gorjanc, and J M Hickey (2021). "AlphaSimR: An R-package for Breeding Program Simulations". In: *G3: Genes—Genomes—Genetics* 11. DOI: 10.1093/g3journal/jkaa017.

Gillespie, John H (1984). "The molecular clock may be an episodic clock." In: *Proceedings of the National Academy of Sciences* 81.24, pp. 8009–8013.

— (1991). *The causes of molecular evolution*. Vol. 2. Oxford University Press, USA.

Gutenkunst, Ryan N et al. (2009). "Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data". In: *PLoS Genetics* 5.10, e1000695.

Haag, Julia, Alexander I Jordan, and Alexandros Stamatakis (2025). "Pandora: A Tool to Estimate Dimensionality Reduction Stability of Genotype Data". In: *Bioinformatics Advances*, vbaf040.

Haller, Benjamin C and Philipp W Messer (2019). "SLiM 3: forward genetic simulations beyond the Wright–Fisher model". In: *Molecular Biology and Evolution* 36.3, pp. 632–637.

Hartfield, Matthew and Sylvain Glémin (2024). "Polygenic selection to a changing optimum under self–fertilisation". In: *PLoS Genetics* 20.7, e1011312.

Hernandez, Ryan D (2008). "A flexible forward simulator for populations subject to selection and demography". In: *Bioinformatics* 24.23, pp. 2786–2787.

Hoskins, Roger A. et al. (2015). "The Release 6 reference sequence of the *Drosophila melanogaster* genome". In: *Genome Research* 25.3, pp. 445–458. DOI: 10.1101/gr.185579.114. eprint: http://genome.cshlp.org/content/25/3/445.full.pdf+html. URL: http://genome.cshlp.org/content/25/3/445.abstract.

Huang, Xin et al. (2023). "dadi-cli: Automated and distributed population genetic model inference from allele frequency spectra". In: *bioRxiv*. DOI: 10.1101/2023.06.15.545182. eprint: https://www.biorxiv.org/content/early/2023/06/16/2023.06.15.545182.full.pdf. URL: https://www.biorxiv.org/content/early/2023/06/16/2023.06.15.545182.

Huber, Christian D, Bernard Y Kim, et al. (2017). "Determining the factors driving selective effects of new nonsynonymous mutations". In: *Proceedings of the National Academy of Sciences* 114.17, pp. 4465–4470.

Huber, Christian D., Arun Durvasula, et al. (July 2018). "Gene expression drives the evolution of dominance". In: *Nature Communications* 9.1. ISSN: 2041-1723. DOI: 10.1038/s41467-018-05281-7. URL: http://dx.doi.org/10.1038/s41467-018-05281-7.

Hudson, Richard R (1983). "Testing the constant-rate neutral allele model with protein sequence data". In: *Evolution*, pp. 203–217.

Hudson, Richard R et al. (1990). "Gene genealogies and the coalescent process". In: *Oxford surveys in evolutionary biology* 7.1, p. 44.

Hudson, Richard R, Kevin Bailey, et al. (1994). "Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster." In: *Genetics* 136.4, pp. 1329–1340.

Hudson, Richard R, Dennis D Boos, and Norman L Kaplan (1992). "A statistical test for detecting geographic subdivision." In: *Molecular biology and evolution* 9.1, pp. 138–151.

Hudson, Richard R and Norman L Kaplan (1995). "Deleterious background selection with recombination." In: *Genetics* 141.4, pp. 1605–1617.

Hunt, Sarah E et al. (Nov. 2018a). "Ensembl variation resources". In: *Database* 2018, bay119. ISSN: 1758-0463. DOI: 10.1093/database/bay119. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/bay119/27329174/bay119.pdf. URL: https://doi.org/10.1093/database/bay119.

— (Nov. 2018b). "Ensembl variation resources". In: *Database* 2018, bay119. ISSN: 1758-0463. DOI: 10.1093/database/bay119. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/bay119/27329174/bay119.pdf. URL: https://doi.org/10.1093/database/bay119.

International HapMap Consortium et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164, p. 851.

Johri, Parul, Brian Charlesworth, and Jeffrey D Jensen (2020). "Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection". In: *Genetics* 215.1, pp. 173–192.

Johri, Parul, Kellen Riall, et al. (2021). "The impact of purifying and background selection on the inference of population history: problems and prospects". In: *Molecular biology and evolution* 38.7, pp. 2986–3003.

Katzman, Sol et al. (2007). "Human genome ultraconserved elements are ultraselected". In: *Science* 317.5840, pp. 915–915.

Kern, Andrew D and Daniel R Schrider (2016). "Discoal: flexible coalescent simulations with selection". In: *Bioinformatics* 32.24, pp. 3839–3841.

— (2018). "diploS/HIC: an updated approach to classifying selective sweeps". In: *G3: Genes, Genomes, Genetics* 8.6, pp. 1959–1970.

Kim, Bernard Y, Christian D Huber, and Kirk E Lohmueller (2017). "Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples". In: *Genetics* 206.1, pp. 345–361.

Kingman, John FC (1982). "On the genealogy of large populations". In: *Journal of applied probability* 19.A, pp. 27–43.

Kyriazis, Christopher C and Kirk E Lohmueller (2024). "Constraining models of dominance for nonsynonymous mutations in the human genome". In: *Plos Genetics* 20.9, e1011198.

Kyriazis, Christopher C, Jacqueline A Robinson, and Kirk E Lohmueller (2023). "Using computational simulations to model deleterious variation and genetic load in natural populations". In: *The American Naturalist* 202.6, pp. 737–752.

Lauterbur, M Elise et al. (2023). "Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations". In: *Elife* 12, RP84874.

Li, Heng and Richard Durbin (2011). "Inference of human population history from individual whole-genome sequences". In: *Nature* 475.7357, p. 493.

Lin, Kao et al. (2011). "Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics". In: *Genetics* 187.1, pp. 229–244.

Liu, Xiaoming and Yun-Xin Fu (2020). "Stairway Plot 2: demographic history inference with folded SNP frequency spectra". In: *Genome biology* 21.1, p. 280.

MacLeod, I M, B J Hayes, and M E Goddard (Sept. 2014). "The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data". In: *Genetics* 198.4, pp. 1671–1684. ISSN: 1943-2631. DOI: `10.1534/genetics.114.168344`. eprint: `https://academic.oup.com/genetics/article-pdf/198/4/1671/42133899/genetics1671.pdf`. URL: `https://doi.org/10.1534/genetics.114.168344`.

McLean, Cory and Gill Bejerano (2008). "Dispensability of mammalian DNA". In: *Genome research* 18.11, pp. 1743–1751.

Morin, Phillip A. et al. (2021). "Reference genome and demographic history of the most endangered marine mammal, the vaquita". In: *Molecular Ecology Resources* 21.4, pp. 1008–1020. DOI: `https://doi.org/10.1111/1755-0998.13284`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13284`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13284`.

Mughal, Mehreen R and Michael DeGiorgio (2019). "Localizing and classifying adaptive targets with trend filtered regression". In: *Molecular biology and evolution* 36.2, pp. 252–270.

Mukai, Terumi et al. (1972). "Mutation rate and dominance of genes affecting viability in Drosophila melanogaster". In: *Genetics* 72.2, pp. 335–355.

Nielsen, Rasmus et al. (2005). "Genomic scans for selective sweeps using SNP data". In: *Genome research* 15.11, pp. 1566–1575.

Peischl, Stephan et al. (2013). "On the accumulation of deleterious mutations during range expansions". In: *Molecular ecology* 22.24, pp. 5972–5982.

Przeworski, Molly (2002). "The signature of positive selection at randomly chosen loci". In: *Genetics* 160.3, pp. 1179–1189.

Przeworski, Molly, Graham Coop, and Jeffrey D Wall (2005). "The signature of positive selection on standing genetic variation". In: *Evolution* 59.11, pp. 2312–2323.

Ragsdale, Aaron P, Alec J Coffman, et al. (2016). "Triallelic population genomics for inferring correlated fitness effects of same site nonsynonymous mutations". In: *Genetics* 203.1, pp. 513–523.

Ragsdale, Aaron P and Simon Gravel (2019). "Models of archaic admixture and recent history from two-locus statistics". In: *PLoS Genetics* 15.6, e1008204.

Ragsdale, Aaron P, Dominic Nelson, et al. (2020). "Lessons learned from bugs in models of human history". In: *The American Journal of Human Genetics* 107.4, pp. 583–588.

Ralph, Peter, Kevin Thornton, and Jerome Kelleher (2020). "Efficiently summarizing relationships in large samples: a general duality between statistics of genealogies and genomes". In: *Genetics* 215.3, pp. 779–797.

Robinson, Jacqueline A et al. (2022). "The critically endangered vaquita is not doomed to extinction by inbreeding depression". In: *Science* 376.6593, pp. 635–639.

Rodrigues, Murillo F, Andrew D Kern, and Peter L Ralph (2024). "Shared evolutionary processes shape landscapes of genomic variation in the great apes". In: *Genetics* 226.4, iyae006.

Sabeti, Pardis C et al. (2002). "Detecting recent positive selection in the human genome from haplotype structure". In: *Nature* 419.6909, pp. 832–837.

Sanchez, Théophile et al. (2021). "Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation". In: *Molecular Ecology Resources* 21.8, pp. 2645–2660.

Santiago, Enrique et al. (2020). "Recent demographic history inferred by high-resolution analysis of linkage disequilibrium". In: *Molecular Biology and Evolution* 37.12, pp. 3642–3653.

Schiffels, Stephan and Ke Wang (2020). "MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent". In: *Statistical Population Genomics*. Ed. by Julien Y. Dutheil. New York, NY: Springer US,

pp. 147–166. 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.

Schrider, Daniel R (2020). "Background selection does not mimic the patterns of genetic diversity produced by selective sweeps". In: *Genetics* 216.2, pp. 499–519.

Schrider, Daniel R and Andrew D Kern (2017). "Soft sweeps are the dominant mode of adaptation in the human genome". In: *Molecular biology and evolution* 34.8, pp. 1863–1877.

Schrider, Daniel R, Fábio K Mendes, et al. (2015). "Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps". In: *Genetics* 200.1, pp. 267–284.

Schrider, Daniel R, Alexander G Shanku, and Andrew D Kern (2016). "Effects of linked selective sweeps on demographic inference and model selection". In: *Genetics* 204.3, pp. 1207–1223.

Schweiger, Regev and Richard Durbin (2023). "Ultrafast genome-wide inference of pairwise coalescence times". In: *Genome Research* 33.7, pp. 1023–1031.

Sheehan, Sara and Yun S Song (2016). "Deep learning for population genetic inference". In: *PLoS Computational Biology* 12.3, e1004845.

Siepel, Adam et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes". In: *Genome research* 15.8, pp. 1034–1050.

Simonsen, Katy L, Gary A Churchill, and Charles F Aquadro (1995). "Properties of statistical tests of neutrality for DNA polymorphism data." In: *Genetics* 141.1, pp. 413–429.

Smith, Chris CR and Andrew D Kern (2023). "disperseNN2: a neural network for estimating dispersal distance from georeferenced polymorphism data". In: *BMC bioinformatics* 24.1, p. 385.

Speidel, Leo et al. (2021). "Inferring population histories for ancient genomes using genome-wide genealogies". In: *Molecular Biology and Evolution* 38.9, pp. 3497–3511.

Tataru, Paula and Thomas Bataillon (2020). "polyDFE: inferring the distribution of fitness effects and properties of beneficial mutations from polymorphism data". In: *Statistical population genomics*, pp. 125–146.

Temple, Seth D, Ryan K Waples, and Sharon R Browning (2024). "Modeling recent positive selection using identity-by-descent segments". In: *The American Journal of Human Genetics* 111.11, pp. 2510–2529.

Terhorst, Jonathan, John A Kamm, and Yun S Song (2017). "Robust and scalable inference of population history from hundreds of unphased whole genomes". In: *Nature Genetics* 49.2, p. 303.

Thornton, Kevin R (2014). "A C++ template library for efficient forward-time population genetic simulation of large populations". In: *Genetics* 198.1, pp. 157–166.

— (2019). "Polygenic adaptation to an environmental shift: temporal dynamics of variation under Gaussian stabilizing selection and additive effects on a single trait". In: *Genetics* 213.4, pp. 1513–1530.

Torres, Raul et al. (2020). "The temporal dynamics of background selection in nonequilibrium populations". In: *Genetics* 214.4, pp. 1019–1030.

Uricchio, Lawrence H and Ryan D Hernandez (2014). "Robust forward simulations of recurrent hitchhiking". In: *Genetics* 197.1, pp. 221–236.

Veeramah, Krishna R et al. (2014). "Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans". In: *Molecular biology and evolution* 31.9, pp. 2267–2282.

Wade, Emma E et al. (2023). "Quantifying the fraction of new mutations that are recessive lethal". In: *Evolution* 77.7, pp. 1539–1549.

Wang, Zhanpeng et al. (2021). "Automatic inference of demographic parameters using generative adversarial networks". In: *Molecular ecology resources* 21.8, pp. 2689–2705.

Williamson, Scott and Maria E Orive (2002). "The genealogy of a sequence subject to purifying selection at multiple sites". In: *Molecular biology and evolution* 19.8, pp. 1376–1384.

Yang, Meng et al. (2022). "TrustGWAS: A full-process workflow for encrypted GWAS using multi-key homomorphic encryption and pseudorandom number perturbation". In: *Cell Systems* 13.9, pp. 752–767.

Zhang, Xinjun, Bernard Kim, Kirk E Lohmueller, et al. (2020). "The impact of recessive deleterious variation on signals of adaptive introgression in human populations". In: *Genetics* 215.3, pp. 799–812.

Zhang, Xinjun, Bernard Kim, Armaan Singh, et al. (2023). "MaLAdapt reveals novel targets of adaptive introgression from Neanderthals and Denisovans in worldwide human populations". In: *Molecular Biology and Evolution* 40.1, msad001.

Zhen, Ying et al. (2021). "Greater strength of selection and higher proportion of beneficial amino acid changing mutations in humans compared with mice and Drosophila melanogaster". In: *Genome Research*

31.1, pp. 110–120. DOI: 10.1101/gr.256636.119. eprint: http://genome.cshlp.org/content/31/1/110.full.pdf+html. URL: http://genome.cshlp.org/content/31/1/110.abstract.

Supplementary Material

Table S1: **Performance of DFE-inference methods in four simulation scenarios**. For each of the three methods, we report the mean absolute error (MAE) of the mean selection coefficient and shape of its inferred DFE. In each simulation scenario, the lowest MAEs are highlighted in bold.

species ID	demography	MAE $E\|s\|$ dadi-cli	MAE $E\|s\|$ grapes	MAE $E\|s\|$ polyDFE	MAE shape dadi-cli	MAE shape grapes	MAE shape polyDFE
HomSap	Constant	**0.0012**	0.0101	0.0023	0.030	0.0086	**0.0068**
HomSap	OOA 5R19	0.014	0.0111	**0.0072**	0.031	0.051	**0.024**
PhoSin	Constant	0.024	0.025	0.024	**0.23**	0.25	0.23
PhoSin	Vaquita2Epoch 1R22	0.024	0.025	**0.024**	**0.18**	0.23	0.21

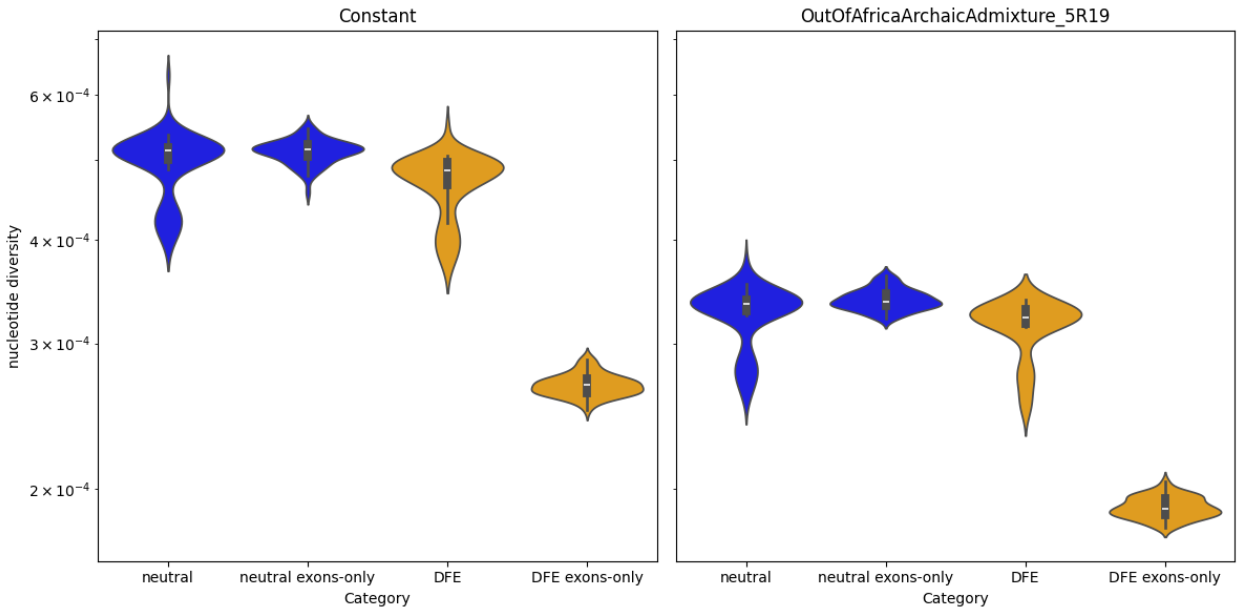


Figure S1: Genetic diversity in human simulations. Nucleotide diversity (π) was calculated for genomes simulated under constant population size (left panel) and under the out-of-Africa demographic model (right panel). Each panel shows four violin plots representing the distribution of π in genomes simulated without selection (blue) and genomes simulated with background selection (orange), calculated from either the entire genome (labeled neutral or DFE), or from exonic regions only (labeled exons-only). Each violin plot summarizes three replicate simulations generated under the relevant scenario.

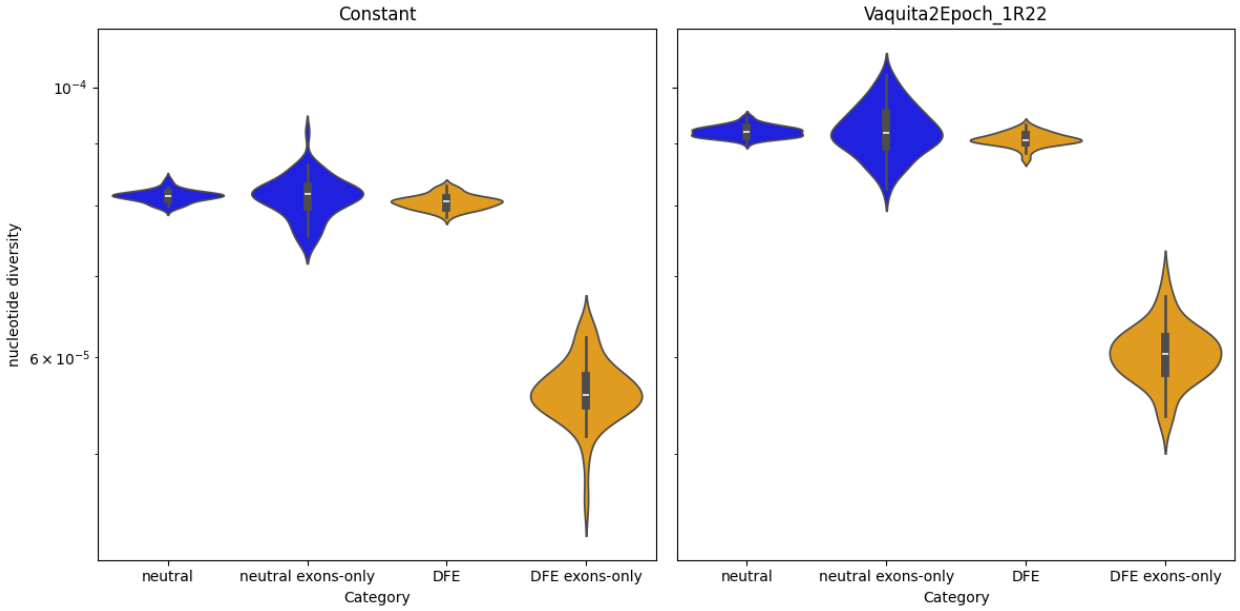


Figure S2: Genetic diversity in vaquita simulations. Nucleotide diversity (π) was calculated for genomes simulated under constant population size (left panel) and under the two-epoch demographic model (right panel). Each panel shows four violin plots representing the distribution of π in genomes simulated without selection (blue) and genomes simulated with background selection (orange), calculated from either the entire genome (labeled neutral or DFE), or from exonic regions only (labeled exons-only). Each violin plot summarizes three replicate simulations generated under the relevant scenario.

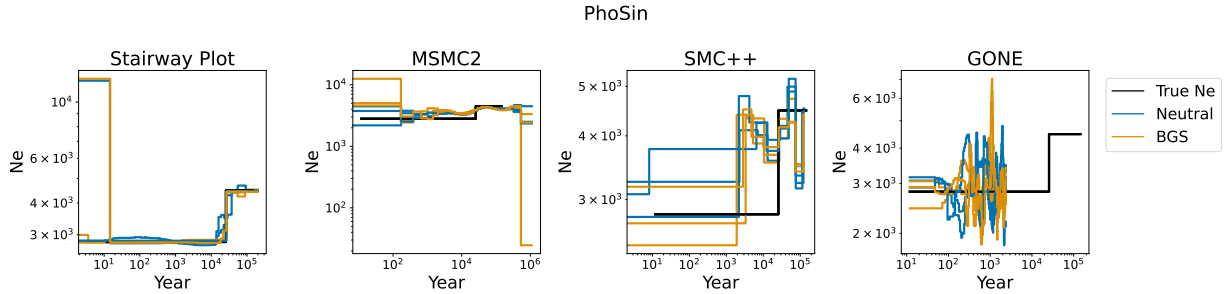


Figure S3: Inference of $N_e(t)$ from vaquita porpoise genomes simulated under a single population model of declining population size with and without purifying selection on exons. Inference is conducted using four methods: `stairwayplot`, `msmc2`, `SMC++` and `GONE`. Each plot depicts the inferred $N_e(t)$ on the three datasets simulated without selection (blue) and the three datasets simulated with the background effects of purifying selection on exons (orange), alongside the true values of N_e used in simulation (black).

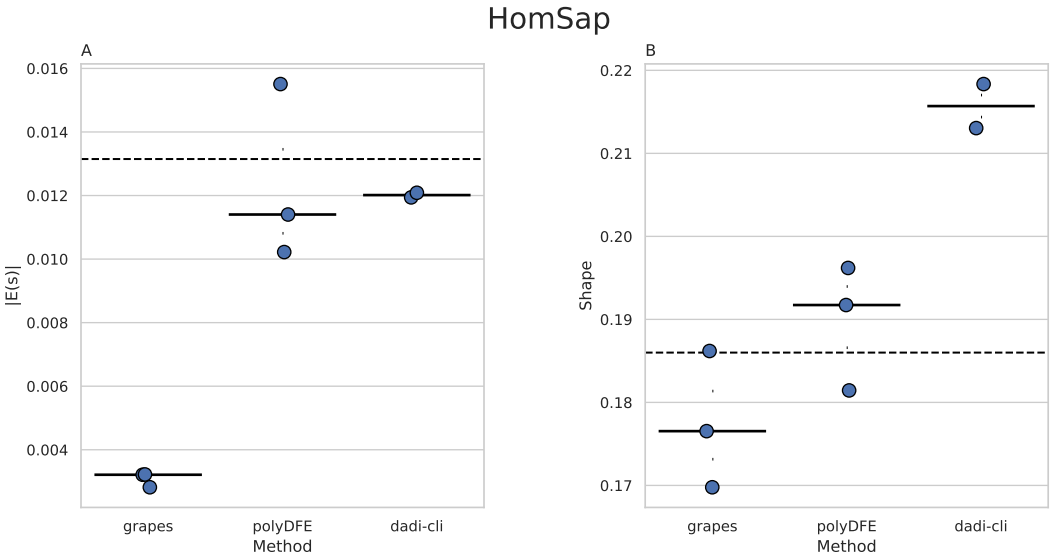


Figure S4: Inferred versus simulated DFEs from human genomes simulated with a simple demographic model with one population with a constant size of $10,000$ and a gamma-distributed DFE acting on exons (see **Methods**). DFE is inferred by three different methods: `GRAPES`, `polyDFE`, and `dadi-cli`. Mean absolute value of selection coefficient ($|E(s)|$, panel A) and shape parameter (panel B) are shown for each DFE inferred from all three simulated datasets with median values marked by horizontal bars and simulated values represented by dashed horizontal lines.

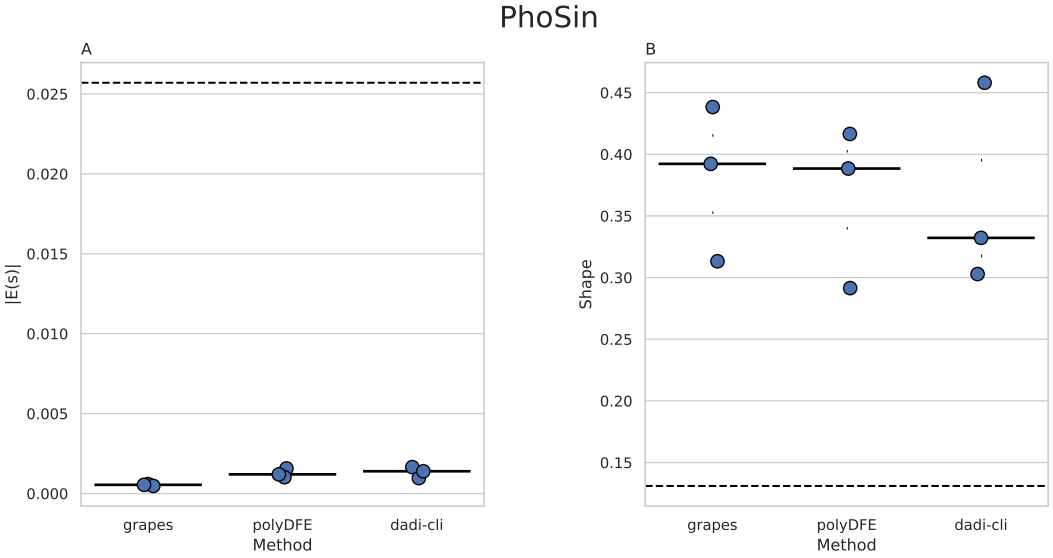


Figure S5: Inferred versus simulated DFEs from vaquita genomes simulated with a simple demographic model with one population with a constant size of $3,500$ and a gamma-distributed DFE acting on exons with a relationship between the selection coefficient (s) and dominance coefficient (h) (see **Methods**). DFE is inferred by three different methods: `GRAPES`, `polyDFE`, and `dadi-cli`. Mean absolute value of selection coefficient ($|E(s)|$, panel A) and shape parameter (panel B) are shown for each DFE inferred from all three simulated datasets with median values marked by horizontal bars and simulated values represented by dashed horizontal lines.

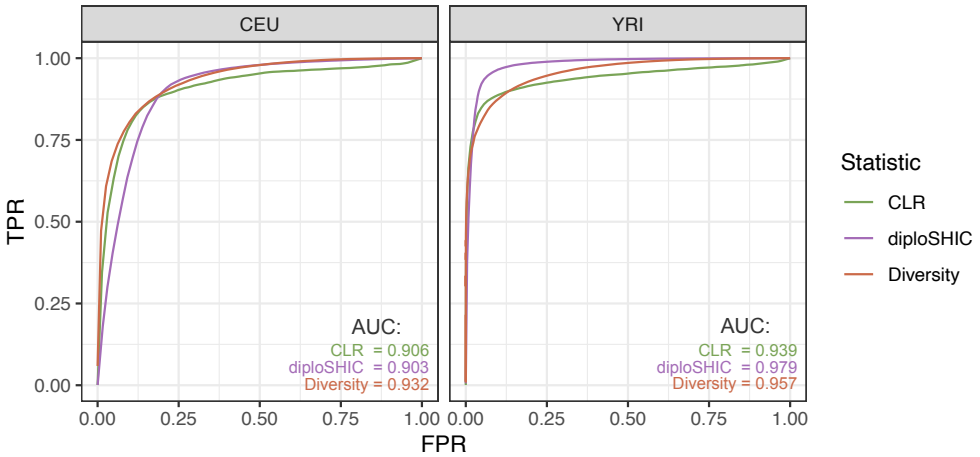


Figure S6: Receiver operating characteristic (ROC) curves for the three methods for detecting sweeps: `sweepfinder2` (labeled CLR), `diploshic`, and reduced diversity (π). True positive rates (TPR) were computed across all datasets simulated with selective sweeps, and false positive rates (FPR) were established by applying the methods to genomes simulated under the same settings but without sweeps (the BGS model; see **Methods**). Area under the curve (AUC) scores are shown for each method.

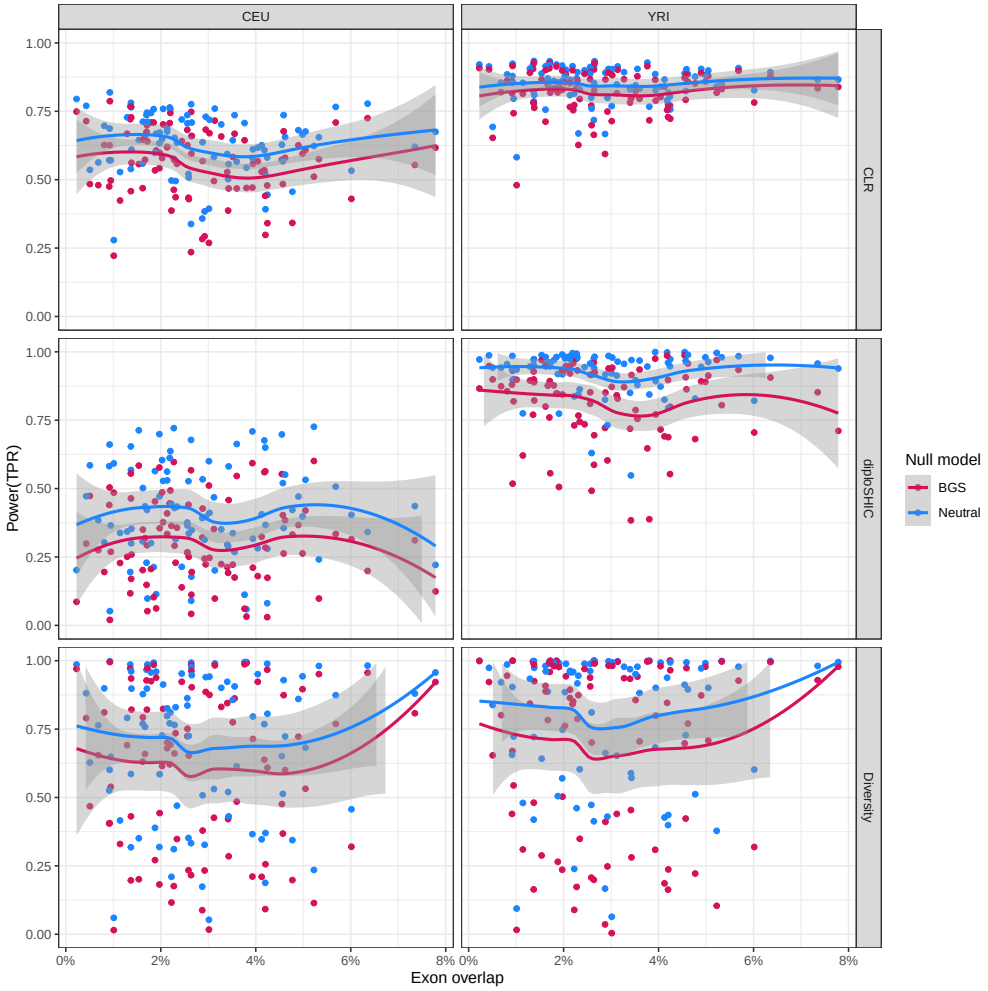


Figure S7: Power to detect selective sweeps as a function of local exon density. This figure shows the same power estimates shown in Figure 5, but with the genomic segments plotted against their exon density (percentage of exonic bases per window) instead of position along chromosome 1. Genomic segments were simulated with sweeps under a three population out-of-Africa model and with background selection from deleterious mutations in exons. Three methods for detecting sweeps were applied to simulated data: `sweepfinder2` (top row– labeled CLR), `diploshic` (middle row), and reduced diversity (π) (bottom row). Power (true positive rate) is shown for these methods for the CEU and YRI samples (left and right respectively). The thresholds of the test statistics were set to obtain a 5% false positive rate under a neutral null model (blue) and a null model with background selection from deleterious mutations in exons (red). Fitted lines represent loess smoothed regressions.