`diplo-locus`: A lightweight toolkit for inference and simulation of time-series genetic data under general diploid selection

Xiaoheng Cheng*[1] and Matthias Steinrücken†[1,2]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA.
[2]Department of Human Genetics, University of Chicago, Chicago IL, USA.

December 1, 2024

Keywords: Time-series genetic data, diploid selection, Hidden Markov Model.

Abstract

Whole-genome time-series allele frequency data are becoming more prevalent as ancient DNA (aDNA) sequences and data from evolve-and-resequence (E&R) experiments are generated at a rapid pace. Such data presents unprecedented opportunities to elucidate the dynamics of genetic variation under selection. However, despite many methods to infer parameters of selection models from allele frequency trajectories available in the literature, few provide user-friendly implementations for large-scale empirical applications. Here, we present `diplo-locus`, an open-source Python package that provides functionality to simulate and perform inference from time-series data under the Wright-Fisher diffusion with general diploid selection. The package includes Python modules as well as command-line tools and is available at: `https://github.com/steinrue/diplo_locus`.

1 Introduction

With the rapid advancements in sequencing technologies, large datasets of temporally stratified population genomic samples are increasingly common. On the one hand, improved techniques for processing ancient DNA (aDNA), that is, DNA from deceased remains, enable the collection of aDNA sequencing data for many samples in humans and other species. On the other hand, for evolution experiments with organisms in the laboratory, it has become increasingly affordable to sequence the genomes of samples taken at multiple time points, bringing forth a fast-growing body of evolve-and-resequence (E&R) datasets. These datasets enable observing changes in the genetic composition of a population across multiple time points, allowing detailed inference about the underlying processes.

One of the most intriguing forces that shape the dynamics of genetic variants is natural or artificial selection, as it reflects how the population adapts to its environment. The Wright-Fisher model and its diffusion approximation are commonly used to model the dynamics of allele frequencies over time, especially when selection is acting, and they have been applied to analyze temporal

*xhcheng@uchicago.edu
†steinrue@uchicago.edu

genetic data (*e.g.*, Bollback et al., 2008; Malaspinas et al., 2012; Mathieson and McVean, 2013; Steinrücken et al., 2014; Steinrücken et al., 2016; He et al., 2020). Nevertheless, despite several existing approaches to infer selection parameters, most analyze the data exclusively under a model of additive (semi-dominant) selection, and are often applied to time-series at specific loci, rather than genome-wide. However, some recent approaches do leverage full genomic data (Mathieson and Terhorst, 2022; Whitehouse and Schrider, 2023).

Here, we present our open-source software package `diplo-locus` that implements the single-locus Wright-Fisher diffusion with piece-wise constant general diploid selection in a panmictic population (see Section S1 in the *Supplemental Material*). It can be used to simulate and analyze temporally sampled genetic data under a Hidden Markov Model (HMM) framework (Bollback et al., 2008). Using efficient Gaussian approximations of the allele frequency dynamics allows for likelihood-based inference of general diploid selection coefficients. With most of the computation vectorized, `diplo-locus` is able to analyze data at a large number of loci simultaneously without compromising efficiency and speed. We believe this lightweight toolkit is a useful addition to the suite of methods for the broader scientific community working with temporal genetic samples.

2 Methods

Our software package `diplo-locus` implements the Wright-Fisher diffusion and the HMM framework (see Section S1 in the *Supplemental Material*) in `Python3`. It requires the `NumPy` (Harris et al., 2020), `SciPy`, `pandas`, and `matplotlib` packages. Once installed, the command-line interface (CLI), `DiploLocus`, can be used to perform analyses in a Unix/Linux environment, or, alternatively, the respective functions can be imported from the `diplo_locus` module in Python. Both, the CLI implementation of this program and the functions of the module can perform likelihood computation and simulation with a wide range of customizable options. A detailed manual and tutorials can be found in the GitHub repository at: `https://github.com/steinrue/diplo_locus`.

2.1 Command-Line Interface `DiploLocus` (CLI)

To compute log-likelihoods, the command `DiploLocus likelihood` reads the genotype data either from a VCF file with a matching sample information file or from a table of allele counts with sampling times. The user can use command-line arguments to designate subsets of loci and filter loci by pooled minor allele frequency. For input as VCF in particular, the user can specify the subset of samples to be considered. If sampled individuals with pseudo-haploid and individuals with diploid genotypes should be analyzed together, the program accepts VCF files that contain both (the command-line options `--force_hap` or `--force_dip` can be used to resolve ambiguities).

In addition to the input files, the user must provide an effective population size (`--Ne`), per-generation recurrent mutation rate(s) (`--u[01|10]`), and the initial allele frequencies (`--initCond`), either as a fixed value or a distribution. For values of per-generation selection coefficients, the program accommodates two fitness parametrizations at a bi-allelic locus with alleles a and A (see Table S1 in the *Supplemental Material*): By relative fitness of heterozygotes s_{aA} and homozygotes s_{AA}, or by selection $s = s_{\mathrm{AA}}$ and dominance coefficients h, with $s_{\mathrm{aA}} = h \cdot s$. The program provides options to compute log-likelihoods on a linear or geometric grid of values for each selection parameter or at a fixed value. The program also allows for selection models with selection coefficients changing over time in a piece-wise constant manner.

Besides computing log-likelihoods on a specified grid, the program can interpolate 1-dimensional log-likelihood surfaces and output on- and off-grid maximum likelihood estimates (MLEs) of the selection coefficients at each locus for increased precision. Furthermore, the program offers the choice to output p-values computed from the likelihood-ratio statistic using a χ^2 distribution with 1 degree of freedom (Self and Liang, 1987).

To simulate temporal samples and allele frequency trajectories, the command `DiploLocus simulate` allows users to simulate an arbitrary number of independent replicates under specified population and selection parameters. The tool can simulate using a model of constant selection throughout time, as well as piece-wise constant selection coefficients. Multiple filtering options based on sample or population frequencies are also provided. As output, the program produces sample allele counts at the specified sampling times, and, optionally, their population allele frequency trajectories. Lastly, the program also allows for convenient visualization of the sample and population frequency trajectories for select replicates.

2.2 Python module `diplo_locus`

The module `diplo_locus.likelihood` contains the class `SelHmm`, which initializes the HMM for computation using specified population and selection parameters as described in Section 2.1. The function `SelHmm.computeLikelihood()` then computes per-locus log-likelihoods based on these parameters for a given number of replicates of temporal samples. The module `diplo_locus.simulate` contains the function `simulateSamples()` to simulate independent replicates based on given population and selection parameters. It returns replicate samples with the corresponding underlying trajectories. Both, analysis and simulation can accommodate piece-wise constant temporal changes in selective pressure.

3 Results

3.1 Statistical Performance

We tested the statistical performance of `diplo-locus` on data simulated using either `diplo-locus`, see Figure 1 and Section S2 in the *Supplemental Material*, or `SLiM4` (Haller and Messer, 2023), see Section S3. We observe from the ROC curves presented in Figure 1A and Figure 1C, that the likelihood-ratio of the estimated MLEs divided by neutrality ($s = 0$) computed using `diplo-locus` can distinguish replicates simulated under selection from neutral replicates. Figure 1B and Figure 1D demonstrate that the MLEs of s_{AA} capture the true value used for the simulations accurately. Additional simulation results in Section S2 and Section S3 of the *Supplemental Material* further demonstrate the power and accuracy of `diplo-locus`.

3.2 Computational Performance

We compared the performance of `diplo-locus` against several other methods to infer selection coefficients from temporal data presented in the literature: `LLS` (Taus et al., 2017), `WFABC` (Foll et al., 2015), and `bmws` (Mathieson and Terhorst, 2022). We simulated 12,000 replicates of temporal data in a variety of scenarios of different length: 160 generations or 4000 generations; additive ($h = 0.5$) or dominance ($h = 1$) selection; and constant selection or time varying selection, see Section S4 of the *Supplemental Material* for a detailed description. Table 1 shows the runtimes for the different

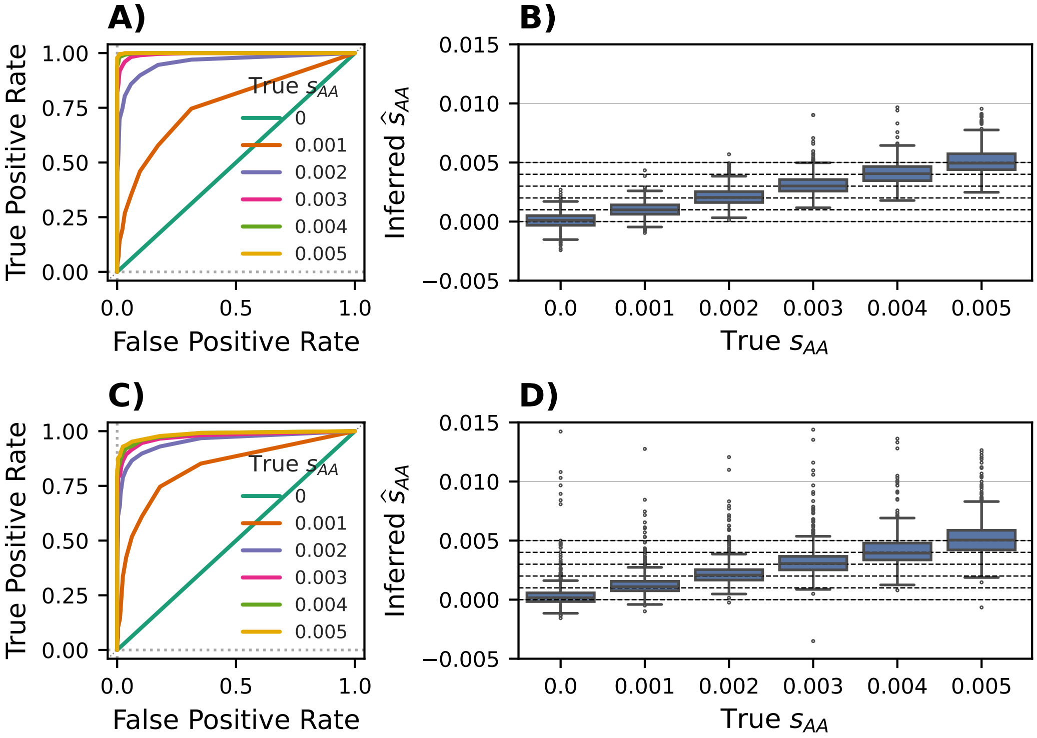

Figure 1: Receiver operating characteristic (ROC) curves demonstrate power to distinguish replicates simulated under **A)** additive selection and **C)** selection with dominance $h = 1$ from neutral replicates using likelihood-ratio tests. Curves based on replicates of 9 temporal samples of 20 diploid individuals, each 500 generations apart, and the diploid population size is $N_e = 10^4$. **B)** & **D)** Boxplots of MLEs $\hat{s}_{\mathrm{AA}}$ for the same set of simulated data used in A) & C) (details in Section S2.1 of the *Supplemental Material*).

# Gener.	h	Var. s	`diplo-locus`	`LLS`	`WFABC`	`bmws`
160	0.5	No	0.054	0.006	296.952	162.193
160	0.5	Yes	0.050	-	-	90.831
160	1.0	No	0.048	0.046	2055.296	-
160	1.0	Yes	0.047	-	-	-
4000	0.5	No	0.077	0.006	691.143	357.558
4000	0.5	Yes	0.068	-	-	379.343
4000	1.0	No	0.068	0.938	248.255	-
4000	1.0	Yes	0.068	-	-	-

Table 1: Comparison of runtimes for different methods. # Gener. is number of generations simulated, h indicates dominance coefficient, and "Var. s" indicates whether s varies over time. Runtimes are given as hours required to analyze all 12,000 replicates in each scenario. We ran `WFABC` and `bmws` for 24 hours and extrapolated the runtimes based on number of completed replicates. Missing value indicates that the method cannot analyze data in respective scenario.

methods to analyze all simulated replicates in each scenario, using a single core of an Intel Ice Lake processor (2.9GHz) with 32GB RAM. Note that the methods from the literature can only be applied to a subset of the scenarios, and thus do not have runtimes reported for all. In addition, Figure S13 and Figure S14 in the *Supplemental Material* show the accuracy of the estimates for the respective methods.

In the scenarios with additive selection, the method `LLS` essentially performs logistic regression, which is highly optimized, and thus the fastest. In all other scenarios, `diplo-locus` is faster or equally fast compared to `LLS`. Both `diplo-locus` and `LLS` are substantially faster than `WFABC` and `bmws`. Due to the vectorized implementation, `diplo-locus` shares the transition matrices for the underlying HMM across replicates, resulting in substantial computational speed-up. Furthermore, Figure S13 and Figure S14 in the *Supplemental Material* show that `diplo-locus` is at least as accurate as the most accurate method among the other three in all scenarios. These results demonstrate that `diplo-locus` is an efficient and flexible method that produces accurate estimates of general diploid selection coefficients in a wide range of scenarios.

3.3 Application to Empirical Data

To showcase the capabilities of `diplo-locus`, we applied the method to three empirical datasets: Temporal data for pigmentation alleles at the ASIP locus in ancient horses (Ludwig et al., 2009), see Section S5.1 in the *Supplemental Material*; Chromosome 2 of ancient humans from Great Britain (GB), dated more recent than 4,500 years before present, extracted from the Allen Ancient DNA Resource (Mallick et al., 2024, v54.1), see Section S5.2; and E&R data for Chromosome 2L of *Drosophila simulans* exposed to a new temperature regime for 60 generations (Barghi et al., 2019), see Section S5.3.

Figure 1A shows the likelihood surface for general diploid selection at the ASIP locus, which is consistent with previous findings (Steinrücken et al., 2014). Figure 1B shows the likelihood surface for the well-characterized SNP `rs4988235` on Chromosome 2 in the GB aDNA human dataset, which regulates expression of the lactase *LCT* gene (Troelsen et al., 2003), with a point MLE of

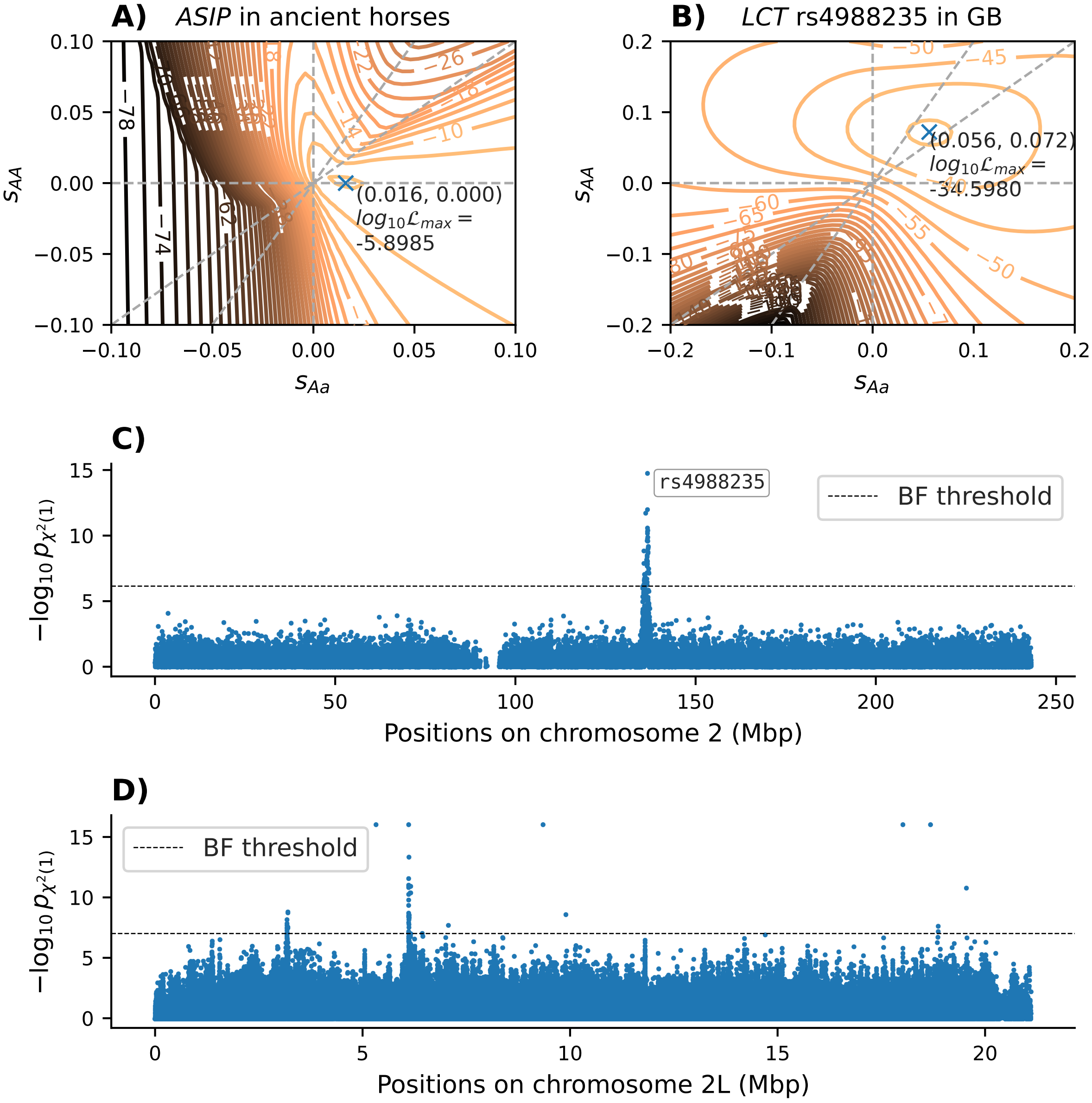

Figure 2: Log-likelihood surface across two-dimensional grid of diploid selection coefficients s_{Aa} and s_{AA}, computed on 101×101 linear grid: **A)** *ASIP* locus in ancient horses from Ludwig et al. (2009) and **B)** SNP `rs4988235` in the *LCT*/*MCM6* locus in ancient human samples from Great Britain (GB). Diagonal dashed gray lines indicate full dominance ($s_{\mathrm{Aa}} = s_{\mathrm{AA}}$) and additivity ($s_{\mathrm{Aa}} = s_{\mathrm{AA}}/2$). **B)** Manhattan plot of p-values for filtered SNPs on Chromosome 2 in the GB aDNA human data. **D)** Manhattan plot of p-values for filtered SNPs on Chromosome 2L from *D. simulans* E&R experiment, combined across 10 biological replicates using Fisher's method.

$(s_{\mathrm{Aa}}, s_{\mathrm{AA}}) = (0.056, 0.072)$. Figure 1C shows a Manhattan plot of the p-values under additive selection for all SNPs on Chromosome 2 in the GB human aDNA dataset, and the SNP `rs4988235` is the most significant signal. Figure 1D shows a Manhattan plot of the p-values under additive selection for Chromosome 2L of the *D. simulans* E&R data, combined across 10 biological replicates using Fisher's method, and Figure S18 and Figure S19 in the *Supplemental Material* show detailed plots of the annotated genes in the vicinity of the two clusters of significant p-values. A detailed analysis of genome-wide results for these datasets is beyond the scope of this manuscript and will be addressed in future work.

As a benchmark, using an Intel Ice Lake processor (2.9 GHz), it takes about 4 hours to compute log-likelihoods for the 520 samples of the GB dataset at the 69,903 genotyped SNPs (after filtering) across a 51-point grid of s_{AA} values. Note that, because the likelihood values are stored in memory before writing output or maximizing, the program might consume more memory than the user's system capacity in cases of a large number of loci or dense parameter grids. Thus, to analyze a larger number of loci, it might be necessary to split the analysis into several parallel batches.

4 Discussion

We present a toolkit developed in Python for analyzing temporal genetic samples under the Wright-Fisher model, with a focus on inferring general diploid selection. Our software can compute likelihoods for time series data, perform likelihood-based inference, and simulate replicates of such data. We assessed the efficiency and accuracy of the inference using our software and showcased the versatility of the CLI `DiploLocus` and the python module `diplo_locus`. We believe this package will be a valuable addition to the toolkit of the population genetics community.

Web Resources

Our software `diplo-locus` is available at: `https://github.com/steinrue/diplo_locus`. Scripts to recreate the simulations and analyses presented in this manuscript are available at: `https://github.com/steinrue/diplo_locus_manuscript_figs`.

Data Availability Statement

The temporal data for the pigmentation alleles in ancient horses is given by Ludwig et al. (2009). The ancient human data is extracted from the Allen Ancient DNA Resource v54.1 described by Mallick et al. (2024), available at: `https://doi.org/10.7910/DVN/FFIDCW` (Dataverse V7). The E&R data for *D. simulans* is described by Barghi et al. (2019), available at: `https://doi.org/10.5061/dryad.rr137kn`.

Acknowledgements

We thank Adam Fine, Jeremy Berg, John Novembre, Maanansa Raghavan, Constanza de la Fuente Castro, and Dylan Sosa for helpful discussions and feedback on this project.

Funder Information

This work is supported by the University of Chicago, a Leakey Foundation Research Grant, and by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under award number R01GM146051.

References

Barghi, N., Tobler, R., Nolte, V., Jakšić, A. M., Mallard, F., Otte, K. A., Dolezal, M., Taus, T., Kofler, R., and Schlötterer, C. Genetic redundancy fuels polygenic adaptation in Drosophila. *PLOS Biology*, 17(2):e3000128, 2019.

Bollback, J. P., York, T. L., and Nielsen, R. Estimation of $2N_es$ from temporal allele frequency data. *Genetics*, 179(1):497–502, 2008.

Foll, M., Shim, H., and Jensen, J. D. WFABC: a wright-fisher abc-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, 15(1):87–98, 2015.

Haller, B. C. and Messer, P. W. SLiM 4: multispecies eco-evolutionary modeling. *The American Naturalist*, 201(5):E000–E000, 2023.

Harris, C. R., Millman, K. J., Walt, S. J., van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with NumPy. *Nature*, 585:357–362, 2020.

He, Z., Dai, X., Beaumont, M., and Yu, F. Estimation of natural selection and allele age from time series allele frequency data using a novel likelihood-based approach. *Genetics*, 216(2):463–480, 2020.

Ludwig, A., Pruvost, M., Reissmann, M., Benecke, N., Brockmann, G. A., Castaños, P., Cieslak, M., Lippold, S., Llorente, L., Malaspinas, A.-S., et al. Coat color variation at the beginning of horse domestication. *Science*, 324(5926):485–485, 2009.

Malaspinas, A.-S., Malaspinas, O., Evans, S. N., and Slatkin, M. Estimating allele age and selection coefficient from time-serial data. *Genetics*, 192(2):599–607, 2012.

Mallick, S., Micco, A., Mah, M., Ringbauer, H., Lazaridis, I., Olalde, I., Patterson, N., and Reich, D. The allen ancient dna resource (AADR) a curated compendium of ancient human genomes. *Scientific Data*, 11(1):182, 2024.

Mathieson, I. and McVean, G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, 193(3):973–984, 2013.

Mathieson, I. and Terhorst, J. Direct detection of natural selection in Bronze Age Britain. *Genome Research*, 32(11-12):2057–2067, 2022.

Self, S. G. and Liang, K.-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82 (398):605–610, 1987.

Steinrücken, M., Bhaskar, A., and Song, Y. S. A novel spectral method for inferring general diploid selection from time series genetic data. *The Annals of Applied Statistics*, 8(4):2203–2222, 2014.

Steinrücken, M., Jewett, E. M., and Song, Y. S. Spectraltdf: transition densities of diffusion processes with time-varying selection parameters, mutation rates and effective population sizes. *Bioinformatics*, 32(5):795–797, 2016.

Taus, T., Futschik, A., and Schlötterer, C. Quantifying selection with pool-seq time series data. *Molecular Biology and Evolution*, 34(11):3023–3034, 2017.

Troelsen, J. T., Olsen, J., Møller, J., and Sjöström, H. An Upstream Polymorphism Associated with Lactase Persistence has Increased Enhancer Activity. *Gastroenterology*, 125(6):1686–1694, 2003.

Whitehouse, L. S. and Schrider, D. R. Timesweeper: accurately identifying selective sweeps using population genomic time series. *Genetics*, 224(3):iyad084, 2023.