PLOS GENETICS

RESEARCH ARTICLE

A novel expectation-maximization approach to infer general diploid selection from time-series genetic data

Adam G. Fine[1,2], **Matthias Steinrücken**[1,3]*

1 Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, United States of America, **2** Graduate Program in Biophysical Sciences, University of Chicago, Chicago, Illinois, United States of America, **3** Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America

* steinrue@uchicago.edu

Citation: Fine AG, Steinrücken M (2025) A novel expectation-maximization approach to infer general diploid selection from time-series genetic data. PLoS Genet 21(7): e1011769. https://doi.org/10.1371/journal.pgen.1011769

Editor: Parul Johri, Arizona State University - Tempe Campus: Arizona State University, UNITED STATES OF AMERICA

Received: December 13, 2024

Accepted: June 11, 2025

Published: July 22, 2025

Data availability statement: All code to generate the results is available at https://github.com/steinrue/EMSel. The ancient DNA data for the analysis was downloaded from the Allen Ancient DNA Resource (AADR) at https://doi.org/10.7910/DVN/FFIDCW (Version 7.0).

Funding: AGF was supported by the Department of Education, Graduate Assistance in Areas of National Need (GAANN), Grant #P200A210054. AGF and MS were supported by the National Institute of General Medical Sciences (NIGMS)

Abstract

Detecting and quantifying the strength of selection is a major objective in population genetics. Since selection acts over multiple generations, many approaches have been developed to detect and quantify selection using genetic data sampled at multiple points in time. Such time-series genetic data is commonly analyzed using Hidden Markov Models, but in most cases, under the assumption of additive selection. However, many examples of genetic variation exhibiting non-additive mechanisms exist, making it critical to develop methods that can characterize selection in more general scenarios. Here, we extend a previously introduced expectation-maximization algorithm for the inference of additive selection coefficients to the case of general diploid selection, in which the heterozygote and homozygote fitness are parameterized independently. We furthermore introduce a framework to identify bespoke modes of diploid selection from given data, a heuristic to account for variable population size, and a procedure for aggregating data across linked loci to increase power and robustness. Using extensive simulation studies, we find that our method accurately and efficiently estimates selection coefficients for different modes of diploid selection across a wide range of scenarios; however, power to classify the mode of selection is low unless selection is very strong. We apply our method to ancient DNA samples from Great Britain in the last 4,450 years and detect evidence for selection in six genomic regions, including the well-characterized LCT locus. Our work is the first genome-wide scan characterizing signals of general diploid selection.

Author summary

Natural selection increases the likelihood that beneficial genetic variants are passed from parent to offspring and thus forms the basis of genetic adaptation to novel environments. Genomic data sampled at multiple timepoints, such as genetic material extracted from

of the National Institutes of Health under award R01GM146051 to MS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

ancient remains (ancient DNA) or data from evolve and resequence experiments, can enable more precise identification of genetic variants subject to selective pressure than contemporary samples alone. However, most methods for identifying genetic variation under selection focus on additive selection, where the fitness of the heterozygote is exactly intermediate between the homozygotes. Leveraging genetic data at multiple timepoints, we develop a method to detect additive and non-additive selection as well as to infer the most likely dominance mechanism. We apply our methods to a dataset of ancient DNA from Great Britain dated less than 4,450 years before present and identify six regions with signals of recent selection, including one at the TFR2 locus that has not been previously reported as a target of selection. Our work enables more accurate quantification of non-additive selection dynamics and can be used to test more complex models of selection.

Introduction

Genetic variation that confers a fitness advantage to an organism over its peers tends to increase in frequency in the population over time until eventual fixation, if it is not lost to genetic drift. This stochastic process of selection ultimately forms the basis of adaptation. Detecting evidence of selection and quantifying its strength is thus a fundamental problem in evolutionary biology, with applications ranging from finding mutations critical to early hominid evolution [1] to predicting tumor growth [2]. In population genetics, many methods to detect signatures of past selective events in contemporary population genomic data have thus been developed [3–5].

However, since selection acts over multiple generations, genetic data observed at several timepoints throughout allows for more accurate quantification of selective processes than present-day samples alone. Recent technological advances have enabled researchers to collect such time-series genetic data genome-wide. One main source of time-series genetic data is ancient DNA (aDNA), that is, genetic material extracted from deceased individuals in humans or other species [6]. Next-generation sequencing has enabled collecting genetic data from large numbers of ancient samples, particularly through the development of techniques such as hybridization enrichment [7]. Another major source of temporal genetic data is experimental evolution studies [8]. Contemporary experimental evolution studies use next-generation sequencing technologies on several biological replicates in evolve and resequence (E&R) experiments to obtain high-quality estimates of temporal allele frequency changes at many loci throughout the genome [9]. These datasets present unprecedented opportunities to detect and characterize the adaptive processes that shape genomic variation [10,11].

Observing the true underlying population allele frequency trajectory as it changes over time would allow for highly accurate characterization of the underlying selective processes. However, in data obtained in practice, genetic variation is often only assessed for a set of individuals sampled at a finite number of time points. Quantifying the strength of selection therefore involves modeling the action of selection, genetic drift, and other population genetic processes on the unobserved trajectory of the population allele frequency, and considering sampled data as imprecise observations of this underlying trajectory.

A commonly used framework for analysis of time-series data that readily accommodates this uncertainty are Hidden Markov Models (HMMs) [12]. In these HMMs, the underlying population allele frequency evolves according to a Markov process, the Wright–Fisher model, and samples are modeled as binomial observations given the underlying population allele

frequency. This HMM framework has been used to estimate a variety of parameters: The additive selection coefficient s [12], the time at which a beneficial mutation arose [13], the effective population size N_e [14], or the rate of sequencing errors [15]. Within this HMM framework, [16] introduced an expectation-maximization (EM) approach, which can be used to estimate additive selection coefficients, as well as migration rates between sub-populations. For reviews of HMM-based approaches to estimate selection coefficients and comparisons between methods in, respectively, aDNA and E&R analyses, see [17] and [18].

Most implementations of the aforementioned HMM approach presented in the literature to date are designed to only detect additive selection, but many examples of genetic variation exhibiting non-additive mechanisms exist. In humans, non-additive targets of selection range from the classical case of the heterozygote advantage conferred by one copy of the sickle-cell allele [19,20] to recent evidence of pervasive dominance in an analysis of data from the UK Biobank [21]. Additionally, stabilizing selection on complex traits, which is believed to be widespread in humans [22], manifests as underdominant selective dynamics at the loci affecting the trait [23–26].

Here, we extend the EM approach in [16] to estimate selection coefficients under a general diploid model, that is, when the fitness values of homozygous and heterozygous genotypes are independently parameterized. The use of an EM method to estimate the selection parameters maximizing the likelihood iteratively allows for better scaling to more than two parameters, compared to grid search-based methods [27]. For example, the use of the EM algorithm allows us to simultaneously estimate diploid selection coefficients and the parameters characterizing the initial frequency of an allele, whereas estimating these parameters using a grid search would be challenging.

We furthermore develop a novel framework for identifying the best-fitting mode of selection for a given temporal dataset. While other methods to estimate general diploid selection coefficients have been presented in the literature [27–32], they do not explicitly address the statistical problem of distinguishing between different modes of selection. Furthermore, none of these methods have been applied to genome-wide data from human populations. To our knowledge, our analysis is the first that characterizes recent general diploid selection in humans from ancient DNA data on a full-genome scale.

The remainder of this article is organized as follows. In Methods, we outline our iterative EM algorithm for efficiently estimating general diploid selection coefficients, as well as the statistical procedure for inferring the most likely mode of diploid selection. In Simulation study, we apply our algorithm and inference framework in a wide range of simulated scenarios to assess its accuracy. We find that our method is generally well-powered to detect selection and estimate its strength, however, power to classify the mode of selection is limited. Moreover, in Inferring effective population size, we introduce a procedure for estimating a constant population size from temporal data. In Ancient DNA dataset from Great Britain, we then perform a genome-wide scan for recent diploid selection in the human genome, using publicly available ancient DNA data from individuals that lived in Great Britain in the last 4,450 years [33, Dataverse v7.0], introduce a procedure to aggregate p-values across linked loci, and discuss six genomic regions that show signals of recent selection. In Coat coloration locus ASIP in domesticated horses, we also apply our method to a locus involved in horse coat coloration [34] to demonstrate the utility of our method when exploring non-additive scenarios. Lastly, we discuss future directions in Discussion. Our method `EMSel` (EM algorithm for detecting Selection) and the scripts to generate the figures in this manuscript are available online at https://github.com/steinrue/EMSel.

Methods

Parameterizing general diploid selection

Throughout this article, we consider selection acting at a given biallelic locus in a diploid population of constant size N_e. The dynamics of the population allele frequency at this locus can be described by the discrete Wright–Fisher model, where we denote by A and a the two alleles at the locus, and by $Y_t \in [0,1]$ and $1 - Y_t$ the random population-level frequency of the A and a allele in generation $t \in \{1, \dots, T\}$, respectively. Suppose that the relative fitness of individuals with genotypes AA, Aa, and aa is $1 + s_2$, $1 + s_1$, and 1, respectively. Given these fitness values, if the frequency of A alleles in the current generation is $Y_t = p$, then the allele frequency in the next generation Y_{t+1} is given by $\frac{1}{2N_e}$ times a binomially distributed random variable with $2N_e$ draws and success probability $p' := p + p(1-p)\big(s_1(1-2p) + s_2 p\big)$ for small s_1 and s_2. This can further be approximated by a normal distribution, where $Y_{t+1} \sim \mathcal{N}\big(p', \frac{1}{2N_e}p(1-p)\big)$, which is commonly referred to as the Wright–Fisher diffusion [35, Ch. 5.3].

We use the term *general diploid selection* for the case of arbitrary $s_1 > -1$ and $s_2 > -1$, that is, the fitness values for the homozygotes and heterozygotes are independently parametrized. Many bespoke modes of selection correspond to constrained diploid selection, where the possible values of s_1 and s_2 are restricted. We consider the following modes of selection:

- **Additive**: $s_1 = \frac{1}{2}s_2$. Relative fitness is proportional to the number of copies of the A allele. Also referred to as **haploid** or **genic** selection.
- **Dominant**: $s_1 = s_2$. Any number of copies of the A allele confers the full fitness effect.
- **Recessive**: $s_1 = 0$. Only individuals homozygous for the A allele have a fitness effect.
- **Overdominance**: $s_1 > \max\{0, s_2\}$. Heterozygous individuals have the highest fitness.
- **Underdominance**: $s_1 < \max\{0, s_2\}$. Heterozygous individuals have the lowest fitness.

Additive, *dominant*, *recessive* selection are one-parameter modes, with $s := s_2 = 2s_1$, $s := s_1 = s_2$, and $s := s_2$ $(s_1 = 0)$, respectively. *Over-* and *underdominance* are two-dimensional subspaces, but we frequently use the version $s := s_1$ $(s_2 = 0)$ as a one-parameter selection mode that describes complete *over-* or *underdominance*. We also refer to the combination of the latter two as *heterozygote difference*. While the one-parameter modes for complete *over-* and *underdominance* are more convenient for simulations, combining them into the one-dimensional mode *heterozygote difference* when inferring parameters avoids statistical artifacts. In Fig 1, we show a diagram of the different modes of selection as sub-spaces of the full two-dimensional parameter space of general diploid selection.

Hidden Markov Model for inferring diploid selection coefficients

Derivation of the EM-HMM algorithm. To derive our novel method for inferring general diploid selection coefficients from sampled time-series genetic data, we extend the method developed in [16] for *additive* selection. If the exact frequency trajectory of the focal allele $p_1, \dots, p_T \in [0,1]$ over T generations is known, then the normal approximation to the Wright-Fisher process can be used to define the likelihood of this trajectory as

$$L^{(C)}_{s_1,s_2}(p_1, \dots, p_T) := \mathbb{P}_{s_1,s_2}\big\{Y_t \in dp_t \,\forall\, 1 \le t' \le T\big\}. \tag{1}$$

Here, we indicate the parameters of interest, the selection coefficients s_1 and s_2, in the subscript, and the superscript (C) indicates the model where the population allele frequency in each generation takes values in the continuous range [0,1]. Without loss of generality, the focal allele is the A allele. Maximizing this likelihood yields a maximum likelihood estimator

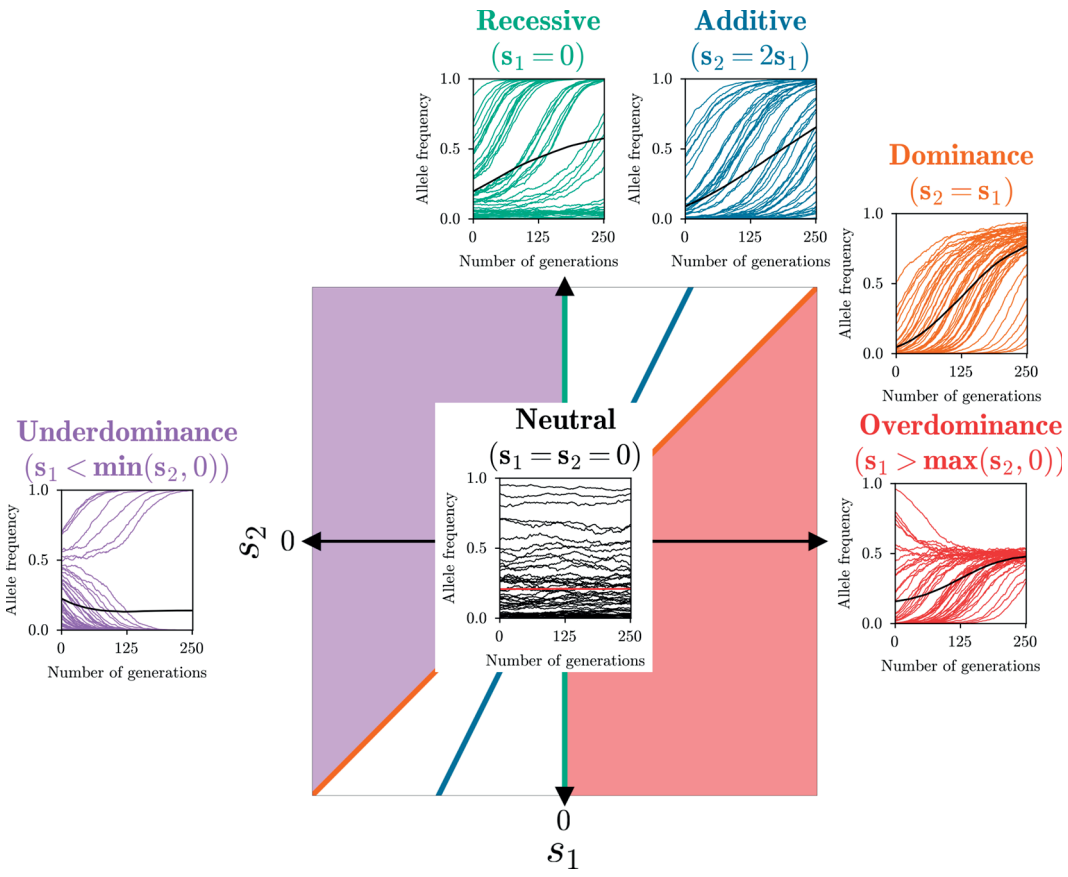

Fig 1. Two-dimensional space of general diploid selection. The five bespoke selection modes we consider are: *Additive*, *dominant*, *recessive*, *over-* and *underdominance*. Colors indicate the sub-space of the respective mode. 50 replicates simulated under each mode, as well as their mean, are plotted to illustrate the characteristic shapes of the trajectories in each mode.

https://doi.org/10.1371/journal.pgen.1011769.g001

(MLE) of the selection coefficients. In the case of *additive* selection, this estimator has been presented in [36]. In Sects S.1.1 and S.1.2 in S1 Text, we revisit this estimator for the *additive* MLE and extend the result to general diploid selection, which yields the estimators

$$\widehat{s_1} = \frac{(p_T - p_1)\sum_t p_t^3 q_t - \sum_t p_t(p_{t+1} - p_t)\sum_t p_t^2 q_t}{\sum_t p_t q_t \sum_t p_t^3 q_t - \left(\sum_t p_t^2 q_t\right)^2}, \text{ and}$$

$$\widehat{s_2} = \frac{\sum_t p_t q_t (1 - 2p_t)\sum_t p_t(p_{t+1} - p_t) - (p_T - p_1)\sum_t p_t^2 q_t (1 - 2p_t)}{\sum_t p_t q_t \sum_t p_t^3 q_t - \left(\sum_t p_t^2 q_t\right)^2},$$

with $q_t := 1 - p_t$.

However, as discussed in Introduction, the data often consists of a given number of individuals sampled from the population at certain points in time, and thus the full trajectory of the population allele frequency is in general not known. To efficiently integrate over this uncertainty, [12] introduced an HMM framework, where the population allele frequency at a given time is the hidden state, evolving according to the Wright-Fisher model, and the sampled genotypes are the observations. Specifically, assume that we have sampled the

population at times $1 \leq t_1, \ldots, t_K \leq T$. At each timepoint, the data $o_{t_i} := (n_{t_i}, a_{t_i})$ consists of the number of haplotypes sampled at this time n_{t_i}, and the number of observed focal alleles a_{t_i}. For convenience, we set $n_t := 0$ and $a_t := 0$ at times where no data is observed, and we denote the random variable associated with sampling the data at t by O_t. In addition, we discretize the population allele frequency into M hidden states to allow efficient computation: $\mathcal{G} := \{g_0 = 0, g_1, \ldots, g_{M-1} = 1\}$, with $g_i \in [0, 1]$, and $g_{i-1} < g_i$. The hidden state in generation t is then the discretized population allele frequency at t, which we denote by $F_t \in \mathcal{G}$.

To apply the standard HMM framework [37,38, Ch. 13.2], we must define initial probabilities $\mathbb{P}_{s_1,s_2}\{F_1 = g_i\}$, transition probabilities, and emission probabilities. For now, we assume that the initial probabilities are given. These can be fixed or estimated, and we provide details on an estimation procedure in Estimation of the initial allele frequency distribution. For the transition probabilities from hidden state g_i to hidden state g_j, we use the normal approximation to the Wright-Fisher process and define

$$\mathbb{P}_{s_1,s_2}\{F_{t+1} = g_j | F_t = g_i\} := \int_{(g_j+g_{j-1})/2}^{(g_j+g_{j+1})/2} \phi\left(x; \mu_i(s_1, s_2), \tfrac{1}{2N_e} g_i(1 - g_i)\right) dx,$$

where $\mu_i(s_1, s_2) := g_i + g_i(1 - g_i)(s_1(1 - 2g_i) + s_2 g_i)$ for general diploid selection and $\phi(x; \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2. Here, we also set $g_{-1} := -\infty$ and $g_M := \infty$, which assigns all the probability mass to transition outside of the interval [0,1] to the respective boundary points. Lastly, the emission probabilities to observe a_t focal alleles in generation t are given by the binomial distribution

$$a_t \sim \text{Bin}(n_t, F_t),$$

with n_t draws and success probability F_t, the respective population allele frequency. Fig 2A depicts a schematic of this HMM.

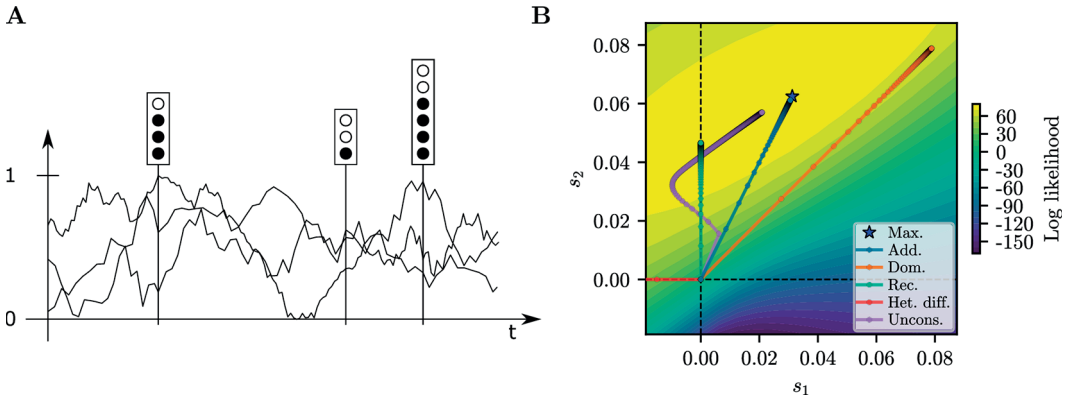

Fig 2. HMM to infer selection. A) Schematic of an HMM. Each "stoplight" represents a haploid sample at the given time, with a certain number of focal alleles. Also plotted are three possible population allele frequency trajectories through the hidden state space. Trajectories with population frequencies closer to the frequencies in the samples are given more weight when computing the expected values in the E-step of the algorithm. B) Log-likelihood surface and path of the EM-HMM optimization under each mode for a given replicate simulated under *additive* selection.

https://doi.org/10.1371/journal.pgen.1011769.g002

The *forward-backward algorithm* [37,38, Ch. 13.2] can then be used to obtain the posterior probability over the hidden states in generation t

$$\gamma_{s_1,s_2}(t,i) := \mathbb{P}_{s_1,s_2}\left\{F_t = g_i \middle| O_{t'} = o_{t'}\ \forall\, 1 \le t' \le T\right\}$$

conditional on the observed data and given selection coefficients s_1 and s_2, and the joint posterior of the hidden states

$$\xi_{s_1,s_2}(t,i,j) := \mathbb{P}_{s_1,s_2}\left\{F_t = g_i, F_{t+1} = g_j \middle| O_{t'} = o_{t'}\ \forall\, 1 \le t' \le T\right\}.$$

Based on these posterior distributions, posterior expectations

$$\mathbb{E}^{(D)}_{s_1,s_2}\left[h(F_t)\middle| O_{t'} = o_{t'}\ \forall\, 1 \le t' \le T\right] = \sum_{i=0}^{M} \gamma_{s_1,s_2}(t,i)h(g_i)$$

and

$$\mathbb{E}^{(D)}_{s_1,s_2}\left[h(F_t,F_{t+1})\middle| O_{t'} = o_{t'}\ \forall\, 1 \le t' \le T\right] = \sum_{i=0}^{M}\sum_{j=0}^{M} \xi_{s_1,s_2}(t,i,j)h(g_i,g_j),$$

can be computed for arbitrary functions $h(\cdot)$ and $h(\cdot,\cdot)$ of the marginal frequency and joint marginal frequencies, respectively. Here, the superscript (D) indicates the use of the model with the discretized allele frequencies.

To find the maximum likelihood estimate (MLE) of the diploid selection coefficients $\widehat{s_1}$ and $\widehat{s_2}$ under this HMM, we use the EM algorithm [37,38, Ch. 13.2], similar to [16] in the *additive* case. We refer to this algorithm as an EM-HMM algorithm. The algorithm starts with an initial parameter estimate $(s_1^{(0)}, s_2^{(0)})$. At iteration k, the algorithm then computes $\gamma_{s_1^{(k)},s_2^{(k)}}(t,i)$ and $\xi_{s_1^{(k)},s_2^{(k)}}(t,i,j)$ (E-step), and updates the parameter estimates by maximizing the conditional log-likelihood using

$$\left(s_1^{(k+1)}, s_2^{(k+1)}\right) = \underset{s_1,s_2\in\mathbb{R}}{\operatorname{argmax}}\ \mathbb{E}^{(D)}_{s_1^{(k)},s_2^{(k)}}\left[\ln L^{(C)}_{s_1,s_2}(F_1,\dots,F_T)\middle| O_t = o_t\ \forall\, 1 \le t \le T\right], \tag{2}$$

(M-step) until the estimates converge. Note that, under the binomial model, the emission probabilities are independent of s_1 and s_2 and do not need to be considered explicitly in this update step. Similar to the derivation of the MLE for a given trajectory from Eq (1), Eq (2) yields

$$s_1^{(k+1)} = \frac{\left(\mathbb{E}[F_T] - \mathbb{E}[F_1]\right)\sum_t \mathbb{E}\left[F_t^2 H_t\right] - \sum_t \mathbb{E}\left[F_t(F_{t+1}-F_t)\right]\sum_t \mathbb{E}\left[F_t H_t\right]}{\sum_t \mathbb{E}\left[H_t\right]\sum_t \mathbb{E}\left[F_t^2 H_t\right] - \left(\sum_t \mathbb{E}\left[F_t H_t\right]\right)^2},$$

$$s_2^{(k+1)} = \frac{\sum_t \mathbb{E}\left[H_t(1-2F_t)\right]\sum_t \mathbb{E}\left[F_t(F_{t+1}-F_t)\right] - \left(\mathbb{E}[F_T] - \mathbb{E}[F_1]\right)\sum_t \mathbb{E}\left[F_t H_t(1-2F_t)\right]}{\sum_t \mathbb{E}\left[H_t\right]\sum_t \mathbb{E}\left[F_t^2 H_t\right] - \left(\sum_t \mathbb{E}\left[F_t H_t\right]\right)^2},$$

where $H_t := F_t(1-F_t)$, and we denote the discretized conditional expectation by

$$\mathbb{E}[\,\cdot\,] = \mathbb{E}^{(D)}_{s_1^{(k)},s_2^{(k)}}\left[\,\cdot\,\middle| O_t = o_t\ \forall\, 1 \le t \le T\right]$$

for brevity. See Sect S.2.2 in S1 Text for details of the derivation.

Note that this approach is not an exact EM algorithm, since it combines the conditional expectation $\mathbb{E}^{(D)}[\,\cdot\,]$ computed under the discretized model with maximization of the likelihood $L^{(C)}(\cdot)$ in the continuous model, similar to [16]. The reason for this hybrid approach is that while the conditional log-likelihood can be maximized in the continuous setting using the diploid generalization of the MLE in [36], the posterior expectations cannot readily be computed. On the other hand, in the discretized model, the posterior expectations can be computed, but the conditional log-likelihood cannot be maximized analytically. The hybrid approach, however, is computationally tractable and yields highly accurate estimates.

Constrained optimization for bespoke selection modes. In addition to estimating unconstrained diploid selection coefficients s_1 and s_2, we also want to estimate selection coefficients in the one-parameter selection modes *additive*, *recessive*, *dominance*, and *heterozygote difference*. To this end, we introduce constraints on s_1 and s_2 using the framework of Lagrange multipliers [39, Ch. 7.5]. Denoting the likelihood by L_{s_1,s_2}, the optimization problem can be formulated as maximizing the conditional log-likelihood $\mathbb{E}[\ln L_{s_1,s_2}]$ in Eq (2) subject to $g(s_1, s_2) = 0$ for an arbitrary function $g(\cdot,\cdot)$, which can be solved by introducing the Lagrangian $\mathcal{L}_{s_1,s_2,\lambda} = \mathbb{E}[\ln L_{s_1,s_2}] - \lambda g(s_1, s_2)$ and solving $\nabla_{s_1,s_2,\lambda}\mathcal{L} = 0$.

All aforementioned one-parameter selection modes of interest can be expressed as a linear constraint, that is, $as_1 - bs_2 = 0$ for suitable $a, b \in \mathbb{R}$. In the Lagrange multiplier formalism, we can thus set $g(s_1, s_2) = as_1 - bs_2$ and solve the constrained optimization problem to obtain the MLE under the respective mode of selection. In Sect S.2.3 of S1 Text, we derive explicit expressions for the update rules for s_1 and s_2 for arbitrary a, b. As an example, Fig 2B shows the iterations of the *additive*, *recessive*, *dominance*, and unconstrained general diploid selection EM-HMM algorithms on a dataset simulated under *additive* selection.

Estimation of the initial allele frequency distribution. The discretized distribution of the initial allele frequency can be fixed for the analysis or can be estimated as well. When estimating it, we fit a beta distribution to the initial frequency, as it is a flexible distribution with only two parameters, denoted by (α, β), which avoids potential over-parametrization. If we assume that the initial distribution does not depend on the selection coefficients, then the discretized version of the EM update rule given in Eq (2) becomes

$$\begin{aligned}\left(\alpha^{(k+1)}, \beta^{(k+1)}\right) &:= \operatorname*{argmax}_{\alpha,\beta\in\mathbb{R}_{>0}} \mathbb{E}\left[\ln L^{(D)}_{\alpha,\beta}(F_1)\right] \\ &= \operatorname*{argmax}_{\alpha,\beta\in\mathbb{R}_{>0}} \sum_{m=0}^{M-1} \gamma_{\theta^{(k)}}(1, m) \ln p(m; \alpha, \beta)\end{aligned} \tag{3}$$

for the parameters (α, β). Here, $L^{(D)}_{\alpha,\beta}(\cdot)$ denotes the likelihood of a discretized beta distribution, the conditional expectation $\mathbb{E}[\cdot]$ is now parameterized by $\theta^{(k)} := \left(\alpha^{(k)}, \beta^{(k)}, s_1^{(k)}, s_2^{(k)}\right)$, and $p(m; \alpha, \beta)$ is the integral of a beta distribution with parameters (α, β) over the m-th discretization interval. At each iteration k of the EM-HMM algorithm, we solve Eq (3) numerically to update α and β alongside the selection coefficients. Since the EM update for the selection coefficients already requires computing $\gamma^{(D)}_{\theta^{(k)}}(t, i)$, the extra step of updating the initial condition comes at minimal computational cost. We observed that estimating the initial distribution does affect the accuracy of the selection coefficient estimation, see Fig J in S1 Text, while providing more flexibility.

Additional implementation details. Analyzing data using our EM-HMM algorithm requires choosing the number of hidden states or discretization intervals M, and where to place the discretization points $\mathcal{G} := \{g_0 = 0, g_1, ..., g_{M-1} = 1\}$. A common choice for the discretization points is to space them equidistantly, that is $g_i = \frac{i}{M-1}$. However, we observed

that this discretization did not perform well when selection was strong, see Sect S.5 in S1 Text for details. We thus decided to use the Chebychev nodes $g_i := \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi i}{2*(M-1)}\right)$ for $0 \leq i \leq M-1$. These nodes are often used for numerical integration, since they mitigate Runge's phenomenon [40, Ch. 4.5]. Intuitively, when using equidistant spacing, the variance of the normal distribution for the transition used in the M-step becomes smaller than the size of the discretization intervals near the boundary, and consequently, the density at the discretization points does not approximate the probability mass in the interval well. Using the Chebychev nodes effectively increases the density of discretization points near the boundary and decreases it in the interior, mitigating this problem.

We analyzed simulated data using different choices of the number of hidden states M in Robustness of selection coefficient estimation. We observed that the inference does not perform well when using fewer than 250 hidden states, but is stable for higher values. To balance accuracy of the EM-HMM algorithm and computational efficiency, we choose $M = 500$ hidden states in all analyses. Unless stated otherwise, we initialize the selection coefficients for the EM-HMM algorithm with $s_1^{(0)} = s_2^{(0)} = 0$ and set the starting parameters for the initial distribution to $\alpha^{(0)} = \beta^{(0)} = 1$, which corresponds to the uniform distribution. For all analyses, we use the convergence criterion that the difference in log-likelihood between iteration k and iteration $k+1$ must be less than 10^{-3}. Additionally, since the EM algorithm can require many iterations to meet this convergence criterion, we tested accelerating the EM using the SQUAREM procedure [41]. We found that, while this approach reduced the total number of iterations required, the total runtime of the algorithm increased due to the additional computational cost per iteration. Since we observed no noticeable increase in accuracy of the estimation, we thus proceeded with the regular EM approach.

We noticed that, especially for simulations under neutrality, the EM-HMM for many replicates would return $s_1^{(0)} = s_2^{(0)} = 0$ as the MLE, but the actual MLE would deviate slightly from 0, with a log-likelihood increase $<10^{-3}$. Nonetheless, this would result in many replicates reporting a log-likelihood ratio of exactly 1, and consequently, p-values around 1 were not well calibrated. To avoid potentially unwanted distortions to the distribution of p-values, we required the EM-HMM to perform at least 5 iterations before stopping, taking the maximum log-likelihood obtained in the first 5 iterations if further iterations do not increase the log-likelihood. In practice, this has minimal effect on replicates simulated under selection, since the EM-HMM requires more than 5 steps to converge in most of these cases.

Distinguishing between modes of selection

P-values for a single alternative. Before describing our approach to the full problem of inferring the mode of selection with multiple alternatives, we first outline the solution to the task of rejecting neutrality in the case of a one-parameter alternative mode of selection. For a given replicate and one-parameter selection mode, we use the EM-HMM algorithm to compute the MLEs $(\hat{s}, \hat{\alpha}, \hat{\beta})$, as well the log-likelihood ll_s for these parameters. Moreover, we compute the MLEs $(\hat{\alpha_0}, \hat{\beta_0})$ and log-likelihood ll_0 under neutrality ($s = 0$). Treating the parameters of the initial distribution as nuisance parameters, we then perform standard likelihood-ratio testing using the likelihood-ratio statistic $D = 2(ll_s - ll_0)$. As the sample size goes to infinity, Wilks' theorem [42] implies that D should be χ^2 distributed with one degree of freedom under the null hypothesis $s = 0$, which we denote by $\chi^2(1)$. However, each replicate consists only of one set of temporal samples, so we are not operating in the asymptotic regime of Wilks' theorem, and have no theoretical guarantee regarding the distribution of D. Nonetheless, in Validation of single-alternative p-values, we show that if we compute p-values using $p = 1 - \mathbb{P}\{\chi^2(1) < D\}$, assuming a $\chi^2(1)$ distribution, then the resulting p-values are

well-calibrated, and thus the asymptotic formula provides a good approximation. P-values computed this way can thus be used to accept or reject the null hypothesis of neutrality for a given dataset, under the given one-parameter alternative.

Multiple alternatives. The mode of selection underlying a given dataset might not be known a priori, so we devised the following strategy that aims at inferring the mode of selection among multiple alternatives from the given data. We found that unconstrained estimation of the MLE $(\hat{s_1}, \hat{s_2})$ in the full two-dimensional parameter space has higher variance than inference in the bespoke constrained modes. Thus, our approach to infer the mode of selection relies solely on the constrained modes of selection. Motivated by the fact that the test statistic D is well-calibrated in the case of single alternative tests for all modes of selection, we propose the following procedure to establish significance and classify significant replicates:

1. For all replicates in the target dataset, obtain the MLE parameters for all constrained modes of selection: *Additive*, *dominance*, *recessive*, and *heterozygote difference*.
2. For each replicate, compute the test statistic

$$\delta := -2\big(ll_0 - \max(ll_{\text{add}}, ll_{\text{dom}}, ll_{\text{rec}}, ll_{\text{het}})\big),$$

   where ll_0 is the likelihood under neutrality, and ll_m is the maximal likelihood for mode m.
3. For each replicate, compute its p-value as $p = 1 - \mathbb{P}\{\chi^2(1) < \delta\}$, and reject neutrality at the desired level of significance.
4. If neutrality is rejected, classify the mode of selection according to

$$\text{argmax}\{ll_{\text{add}}, ll_{\text{dom}}, ll_{\text{rec}}, ll_{\text{het}}\}.$$

   In the case that the *heterozygote difference* mode is the most likely mode, classify as *overdominant* if $s_1 > 0$ and as *underdominant* if $s_1 < 0$.

We found that the p-values computed from the statistic δ using a $\chi^2(1)$ distribution with 1 degree of freedom are not as well calibrated as the p-values in the single alternative tests, and can lead to slightly anti-conservative p-values (see Validation of single-alternative p-values) in the tail of the distribution. This is not unexpected, as some p-values from the well-calibrated single-alternative D statistic are replaced by higher-likelihood estimates from additional modes of selection. However, using a $\chi^2(2)$ distribution resulted in an even poorer fit.

Additionally, we explored fitting parametrized distributions to parametric bootstrap simulations of the statistic δ. In Fig N in S1 Text, we compare p-values computed using the $\chi^2(1)$ and $\chi^2(2)$ distribution to p-values computed using parameterized distributions, such as the generalized gamma distribution, that were fit to bootstrap simulations with 10,000 replicates in a scenario related to our data analysis in Ancient DNA dataset from Great Britain. While the bootstrap procedure achieved better calibration in this scenario, the procedure required extensive simulations specific to a given scenario to determine the requisite parameters. We thus recommend using the $\chi^2(1)$ distribution, since it is a fast and flexible procedure that leads to reasonably well-calibrated p-values, even in the tail of the distribution, and is partly motivated by theory. In practice, we do recommend simulating data in the specific scenario of interest to confirm that the p-values are well calibrated.

Results

Simulation study

To assess the performance of our EM-HMM algorithm to estimate selection coefficients and infer the correct mode of selection across a variety of selection regimes, we simulated datasets under several combinations of parameters and exhibit the accuracy of the inference.

Simulation parameters. We simulated population allele frequency trajectories under the discrete Wright-Fisher model for a given number of generations and sampled a certain number of haplotypes at given times from a binomial distribution with success probability given by the respective population allele frequency. Specifically, we simulated datasets under neutrality ($s_1 = s_2 = 0$), each one-parameter selection mode (*additive*, *recessive*, *dominant*), where we set $s = s_2$, and complete *over-* or *underdominance* with $s = s_1$ and $s_2 = 0$. We simulated using selection coefficients $s \in \{0.005, 0.01, 0.025, 0.05\}$ and the length of the simulated trajectory in generations $T \in \{101, 251, 1{,}001\}$. We considered two different types of initial conditions: 1) The population allele frequency is initialized at a fixed value $p \in \{0.005, 0.25\}$ for each replicate, and 2) the initial frequency of each replicate is drawn from the set $i \in \{\frac{1}{2N_e}, \ldots, \frac{2N_e-1}{2N_e}\}$ with probability proportional to $\frac{1}{i}$. The latter is the stationary distribution for neutral segregating sites under the Poisson Random Field model [43], and thus corresponds to selection from standing variation. We conditioned the simulated data on the focal allele A not being lost or fixed in the samples, depending on the mode of selection. Specifically, for *additive*, *dominant*, and *recessive* selection, we conditioned on A not being fixed at the first generation and not being lost at the last generation; for *overdominance*, the simulation is conditioned on A segregating in the last generation; for *underdominance*, we conditioned on A segregating in the first generation.

For all simulations, the population size was set to $N_e = 10{,}000$. the sampling scheme consisted of $K = 11$ equidistantly-spaced points in time, where the first and last point align with the start and the end of the simulated trajectory, respectively. At each time, we sampled $n_{t_k} = 50$ haploids. For each combination of parameters, we simulated 10,000 replicates under neutrality, and 1,000 replicates under each mode of selection and strength of selection. For each replicate, we performed inference using the same N_e that was used for the simulations under the four bespoke selection modes: *additive*, *recessive*, *dominant*, and *heterozygote difference*.

In Estimating selection coefficients, Validation of single-alternative p-values, and Validation of selection mode inference, we present results for the simulations with 251 generations and initial allele frequency fixed to $p = 0.25$. The results for other combinations of number of generations and initial allele frequency distribution can be found in Sect S.3 of S1 Text. In general, the results are similar, with the exception of extreme cases: For *recessive* or *underdominance* with low initial frequencies, the selected allele is often lost, resulting in inaccurate estimates.

Estimating selection coefficients. In Fig 3A, we present the distribution of the estimated selection coefficient $\hat{s}$ across replicates simulated under the *additive*, *recessive*, *dominant*, *overdominant*, and *underdominant* modes. Estimates of the selection coefficients are in general accurate and unbiased. For intermediate-to-strong selection, *overdominance* has the highest variance in $\hat{s}$ estimates. The likely explanation is that the *overdominance* replicates reach stationarity ($p = 0.5$) quickly, resulting in fewer informative samples, and increased variance of estimates. In Fig I of S1 Text, we plot the selection coefficients estimated using the unconstrained EM for all modes of selection, that is, estimating both s_1 and s_2. *Additive*, *recessive*, *dominant*, and *overdominant* selection are well-estimated by the unconstrained

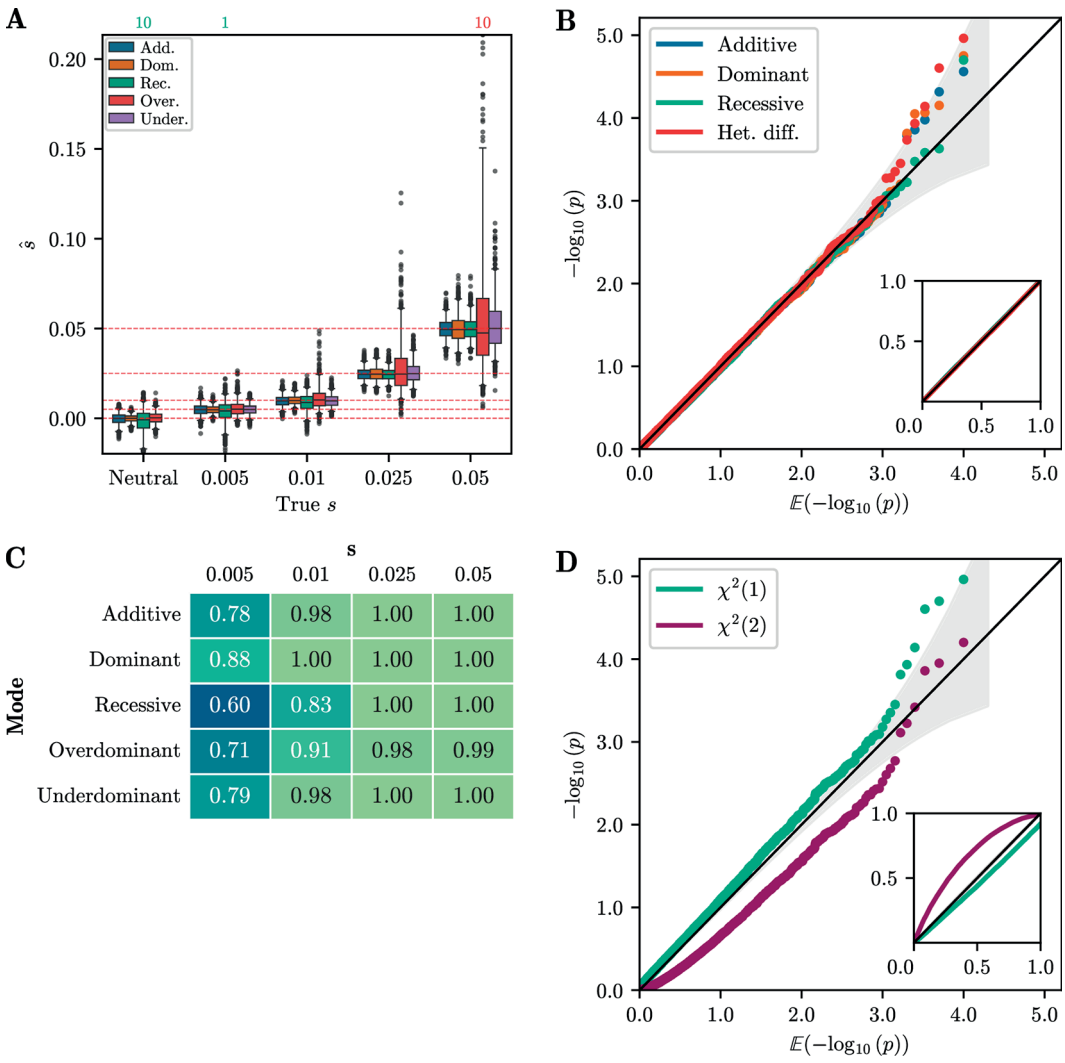

Fig 3. Accuracy of inference from simulated data. A) Boxplot of $\hat{s}$ for 1,000 replicates simulated under each selection mode. Whiskers extend to 2.5% and 97.5%-tiles. Number of estimates outside plotting range indicated above the plot. Estimates are generally unbiased and have low variance, with the exception of *overdominance*. B) Q-Q plot of $-\log_{10}(p)$ against $\mathbb{E}[-\log_{10}(p)]$ of single-alternative tests for neutral replicates. Inset shows same plot for raw values. The p-values are well-calibrated under all modes of selection. C) Table of AUC values based on likelihood-ratios for each selection mode and selection strength simulated. For s>0.01, AUC values are near 1, indicating perfect discrimination between neutral and non-neutral replicates. D) Q-Q plot of $-\log_{10}(p)$ against $\mathbb{E}[-\log_{10}(p)]$ for the δ statistic under the $\chi^2(1)$ and $\chi^2(2)$ distributions. The $\chi^2(1)$ distribution is better calibrated, although it is slightly anti-conservative in the tail. For all simulations, the number of generations is $T = 251$ and the initial condition is fixed at frequency $p = 0.25$.

https://doi.org/10.1371/journal.pgen.1011769.g003

EM, whereas the estimates for *underdominance* show more variability. We thus recommend exploring the bespoke one-dimensional modes first.

Validation of single-alternative p-values. Next, we report the single-alternative p-values obtained using the $\chi^2(1)$ distribution, as described in P-values for a single alternative, for all four modes: *additive*, *recessive*, *dominant*, and *heterozygote difference*. Fig 3B shows a Q-Q plot, where we plot the empirical p-values against their expected value for replicates simulated under neutrality. For all selection modes, the points follow the line $y = x$ closely, indicating

that the p-values are well-calibrated, and that the distribution of the likelihood-ratio statistic D is well-approximated by a $\chi^2(1)$ distribution.

Additionally, we report the performance of the single-alternative tests in terms of the area under the curve (AUC) of a receiver-operator characteristic curve. An AUC value of 0.5 indicates no power to distinguish neutral replicates from non-neutral replicates, whereas an AUC value of 1 indicates perfect discrimination. Fig 3C reports the AUC values for each simulated selection mode over the range of selection parameters s. We observe that in these scenarios, the method is well-powered to reject the null hypothesis for $s \geq 0.01$. AUC values are lower in the case of *recessive* selection, since 251 generations can be insufficient for replicates at low frequency to escape the drift-dominated regime.

Validation of selection mode inference. We furthermore computed p-values for the simulated datasets in the case of multiple alternatives, following the procedure outlined in Multiple alternatives. A Q-Q plot of the p-values against their expected values is shown in Fig 3D. In this scenario, the $\chi^2(1)$ distribution provides a good fit for the δ statistic, although it is slightly anti-conservative in the tail of the distribution. However, in other scenarios (see Sect S.3 in S1 Text), the $\chi^2(1)$ distribution can be slightly *over*conservative. Since the different alternative selection modes are all embedded in a two-dimensional parameter space, Fig 3D also shows p-values computed from the δ statistic using a $\chi^2(2)$ distribution. However, we observe that these p-values are poorly calibrated, and thus recommend using the $\chi^2(1)$ distribution.

We then tested the ability of our approach to correctly identify the mode of selection. Fig 4 shows a confusion matrix resulting from inferring the mode of selection with a p-value significance threshold of 0.05. Each row of the confusion matrix represents all replicates simulated under a particular mode of selection. The numbers in the corresponding column indicate the fraction of replicates in which a particular mode is inferred. Values on the diagonal reflect identification of the correct mode, whereas off-diagonal values reflect replicates where the mode is not correctly inferred. For $s = 0$, neutrality is rejected for 6.3% of replicates, indicating that the $\chi^2(1)$ distribution is anti-conservative. For $s = 0.05$, the correct model is inferred for a majority of replicates for all modes of selection. For weaker selection, $s = 0.01$, only *dominant* and *overdominant* selection are inferred for a majority of replicates. However, neutrality is rejected for over 50% of the replicates for all modes of selection, indicating that power to reject neutrality exists, but accuracy to correctly infer the mode of selection is more limited.

Lastly, we investigate accuracy of the MLE of the selection parameters, given that the mode was correctly classified. Fig 5 shows the distribution of $\hat{s}$ estimates for replicates that were correctly classified as *additive*, *dominant*, or *recessive* selection, as well as $\hat{s}$ for neutral replicates that were classified *incorrectly*. For larger s, correctly classified replicates are more tightly clustered around the true value, with far fewer outliers compared to the distribution using all replicates. For lower values of s and neutrality, the $\hat{s}$ estimates for replicates classified as non-neutral are biased. This is primarily due to a "winner's curse" effect, where neutrality is only rejected for replicates with extreme MLEs.

Robustness of selection coefficient estimation. Next, we investigated how estimation accuracy is affected if we vary parameters that were fixed in the previous simulation study. We varied the number of times a sample is taken $K \in \{2, 5, 11, 35, 101\}$, the number of haploid samples at each timepoint $n_{t_k} = \{6, 20, 50, 100, 200\}$, the effective population size $N_e \in \{10^2, 10^3, 10^4, 10^5, 10^6\}$, and the number of hidden states in the HMM $M \in \{100, 250, 500, 1{,}000, 2{,}000\}$. We varied one parameter at a time, and fixed the others to $K = 11$, $n_{t_k} = 50$, $N_e = 10{,}000$, and $M = 500$. We fixed the selection coefficient s to 0.025 for all simulations, set the total number of generations to $T = 251$, and set the initial distribution

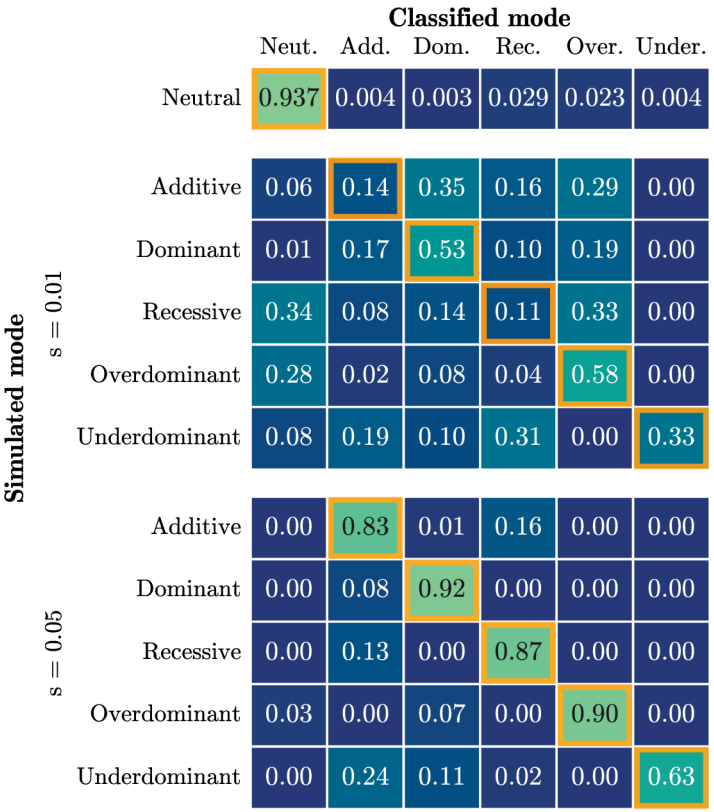

Fig 4. Confusion matrix when inferring the mode of selection. A p-value threshold of 0.05 is used to reject neutrality. Cells with an orange border represent correct classification. The correct model is inferred for the majority of replicates for $s = 0.05$ for all modes of selection. For all simulations, the number of generations is $T = 251$ and the initial condition is fixed at frequency $p = 0.25$.

https://doi.org/10.1371/journal.pgen.1011769.g004

to a fixed frequency at $p = 0.25$. Boxplots of the estimated selection coefficients for *additive*, *recessive*, and *dominant* selection are shown for each scenario in Fig 6. Similar to [16] in the *additive* case, we find that the accuracy does not depend strongly on the sampling scheme except for very low values, such as sampling 6 haplotypes per timepoint or only sampling at the beginning and end of the trajectory. However, the boxplots are tighter for larger sample sizes and more sampling times. This implies that both variance due to finite sampling and variance in the true allele frequency contribute to uncertainty of the selection coefficients estimates. Similarly, the accuracy of the estimates are comparable for 250, 500, 1,000, and 2,000 hidden states. However, for 100 hidden states, the estimates are strongly biased downward. We thus recommend using $M = 500$ to balance accuracy with computational efficiency.

To investigate the accuracy of the estimates when the mode of selection is incorrect, we re-analyzed a subset of the data under scenarios in which we misspecify the mode. We consider two cases: (1) Simulating under *additive*, *dominant*, *recessive*, *over-*, or *underdominant* selection and analyzing under *additive* selection, and (2) simulating under *additive* selection and analyzing under *additive*, *dominant*, *recessive*, or *heterozygote difference*. Fig 7A and 7B depict boxplots of the MLEs $\hat{s}$ under both of these misspecification scenarios.

When simulating under each mode of selection and analyzing under the *additive* EM-HMM, the estimates of $\hat{s}$ are biased for all modes, most strongly for *overdominant*

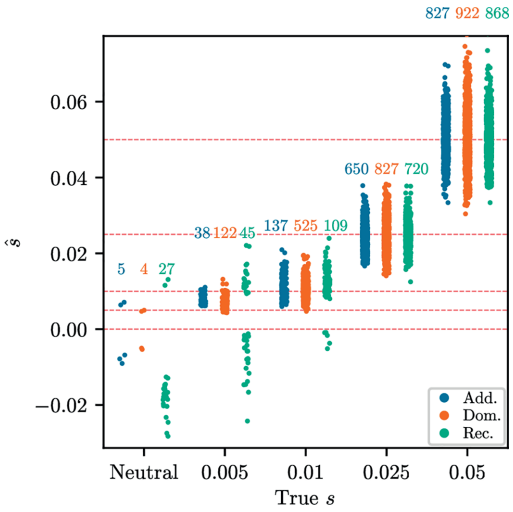

Fig 5. Strip plots of inferred $\hat{s}$ against true s. Conditioned on the correct model being chosen among multiple alternatives. In the neutral case, the plot shows the neutral replicates classified into certain modes. For neutrality, the strip plot is for 1,000 replicates randomly chosen from 10,000 simulated. The number over each strip indicates the number of replicates correctly classified out of the 1,000. For $s = 0.005$ and neutrality, the inferred values are strongly biased, due to a "winner's curse" effect. For all simulations, the number of generations is $T = 251$ and the initial condition is fixed at frequency $p = 0.25$.

https://doi.org/10.1371/journal.pgen.1011769.g005

selection at high s. Alleles undergoing strong *overdominant* selection quickly reach a stationary frequency of $p = 0.5$. Since the alleles do not continue to increase in frequency, the *additive* EM-HMM thus underestimates the true selection strength. In the case when we simulate under *additive* selection and analyze under each selection mode, the $\hat{s}$ estimates are not strongly biased for *recessive* and *dominant*, but reverse sign for *heterozygote difference* at high selection strengths. In this case, the allele frequency increases beyond $p = 0.5$, so the *heterozygote difference* EM-HMM changes from *overdominance* to *underdominant* selection. Overall, the biases in the estimates $\hat{s}$ indicate that using the incorrect one-parameter EM-HMM provides inaccurate estimates of s.

We furthermore investigated the power to reject or accept neutrality in the single-alternative testing framework described in P-values for a single alternative under each of these misspecification scenarios. The resulting AUCs are presented in Fig 7C and 7D. For the case of analyzing other modes under *additive* selection, AUC values are slightly lower than the values in Fig 3C, but still substantially greater than 0.5. For the *additive* datasets analyzed using the incorrect selection modes, the AUC values do not differ substantially between different modes. In both misspecification scenarios, the one-parameter *additive* EM-HMM has sufficient power to accurately reject or fail to reject neutrality for moderate-to-high selection coefficients. Thus, if the goal is only to identify non-neutral evolution, it can be sufficient to analyze given data using the *additive* EM-HMM only, but for accurate characterization of the selection coefficients, the correct mode needs to be used.

Inferring effective population size. We next explored applying our method to infer the effective population size N_e of the underlying population. The approach to derive the update rules for the EM-HMM algorithm provided in Eq (2) and Eq (3) does not readily yield an update rule for N_e. Thus, we instead use a grid-based approach to estimate it. Specifically, we compute the likelihood of observing a given replicate under the neutral HMM with $s_1 = s_2 = 0$ and a given population size N_e. To combine power across replicates or loci, we compute the

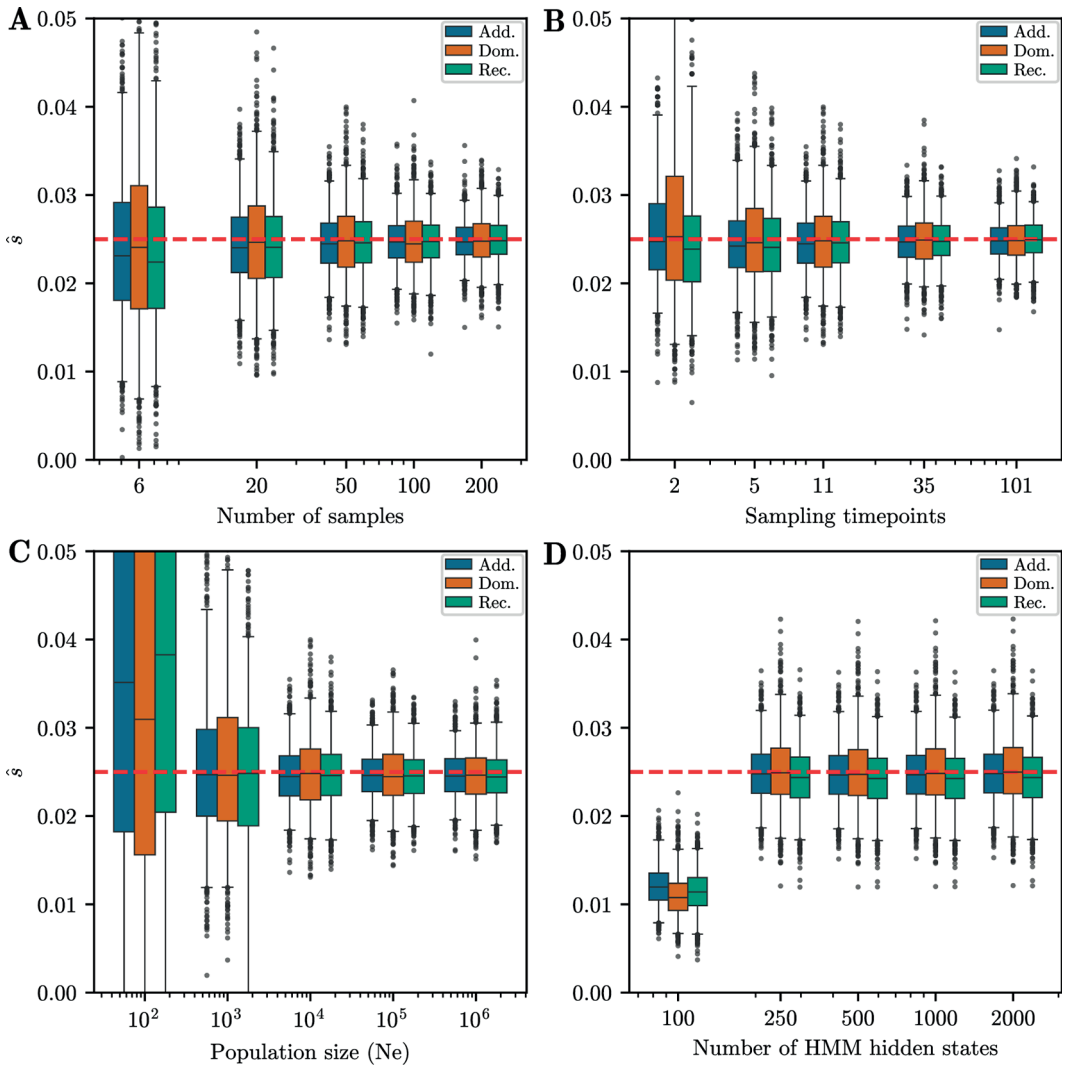

Fig 6. Boxplots for different parameter ranges. Whiskers extend to 2.5% and 97.5%-tiles. We vary: A) Number of samples taken at each timepoint, B) Number of timepoints sampled, C) Effective population size, D) Number of hidden states in the HMM. Except for the lowest parameters, estimates of the selection coefficient are unbiased. Width of boxplots decreases as the number of samples and sampling timepoints increases. For all simulations, the number of generations is $T = 251$ and the initial condition is fixed at frequency $p = 0.25$. When not varied, parameters used are $K = 11$ sampling times, $n_{t_k} = 50$ samples at each timepoint, $N_e = 10{,}000$, and $M = 500$ hidden states.

https://doi.org/10.1371/journal.pgen.1011769.g006

sum of the log-likelihoods over all replicates on a grid of N_e values, then interpolate between these grid values and use the value of N_e that maximizes this composite likelihood surface as our estimate of N_e.

We noticed that when simulating data with large values of N_e, the resulting likelihood surface would often be very flat, making estimation challenging. To counteract this, we condition our likelihoods on observing at least one polymorphic sample at any timepoint, by dividing the likelihood by $\left(1 - \mathbb{P}\{\text{no focal alleles observed}\} - \mathbb{P}\{\text{no non-focal alleles observed}\}\right)$. This penalizes high values of N_e and results in more peaked likelihood surfaces. Fig Q of S1 Text shows several example composite likelihood surfaces. Even after this conditioning procedure, the surfaces are still fairly flat, but they do allow for determining a clear maximum.

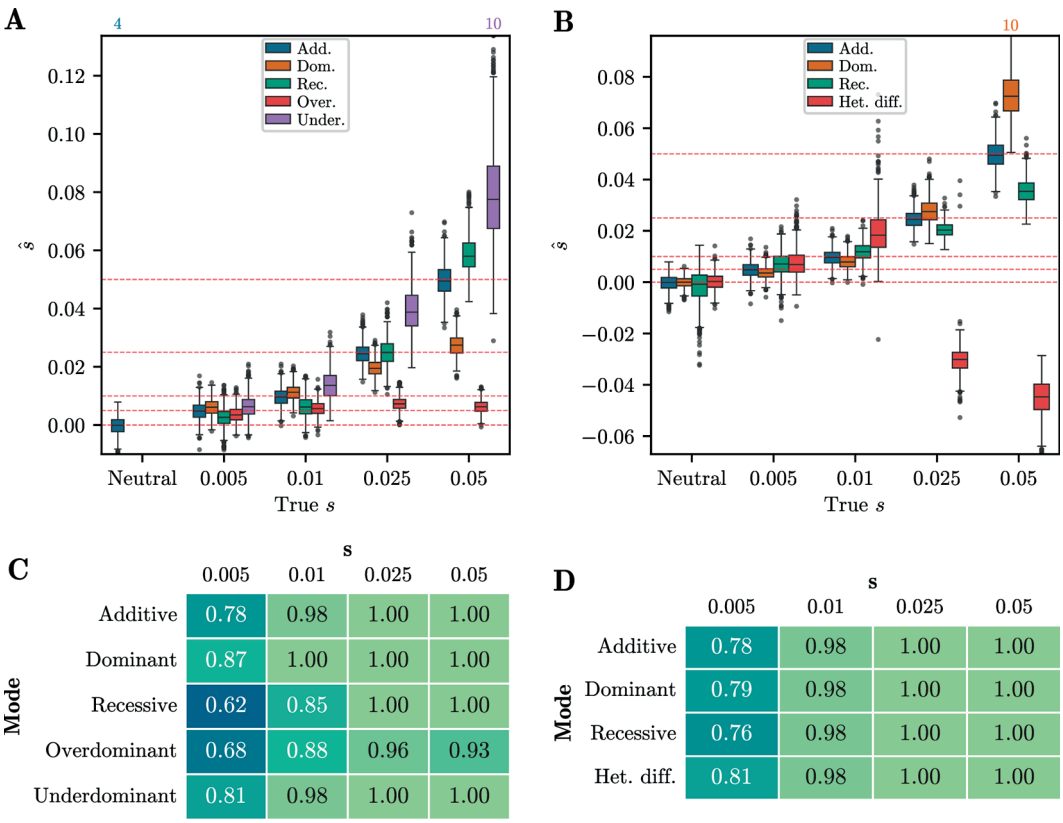

Fig 7. Inference using incorrect selection mode. A) Boxplots of $\hat{s}$ for 1,000 replicates simulated under each selection mode and analyzed using *additive* selection. The estimates are mostly inaccurate. B) Boxplots of $\hat{s}$ for 1,000 replicates simulated under *additive* selection and analyzed under each selection mode. Again, the estimates are not accurate if the mode is incorrect. For both sets of boxplots, whiskers extend to 2.5% and 97.5%-tiles. C) Table of AUC values for data simulated under each selection mode using likelihood-ratios obtained assuming *additive* selection. Values are lower compared to analyzing under the correct mode of selection, but still show substantial power to reject neutrality. D) Table of AUC values for data simulated under *additive* selection using likelihood-ratios obtained assuming each possible selection mode. The values are very similar across selection modes. For all simulations, the number of generations is $T = 251$ and the initial condition is fixed at frequency $p = 0.25$.

https://doi.org/10.1371/journal.pgen.1011769.g007

Furthermore, we found that the N_e estimates were most accurate when the initial condition for computing the likelihood is fixed as the uniform distribution, rather than estimated as in Estimation of the initial allele frequency distribution. A possible explanation could be that the procedure to estimate the initial condition is affected by the choice of N_e, which in turn could bias the estimates.

We simulated 25 batches of 10,000 neutral replicates under the scenario with 251 generations, initial frequency fixed at $p = 0.25$, and the sampling scheme used in Simulation study for $N_e \in \{2{,}500, 10{,}000, 40{,}000\}$. We estimated N_e for each batch using the above procedure. Fig 8 shows boxplots of the inferred N_e for each batch. The N_e values estimated from the grid-based procedure are tightly clustered around the true value, although slightly biased upward in all cases. More powerful approaches to estimate N_e exist, for example, using IBD segments in contemporary data [44]. However, we believe that when analyzing time-series genetic data, the performance of the grid-based HMM procedure is acceptable and yields the most appropriate estimate of N_e to use in downstream analyses.

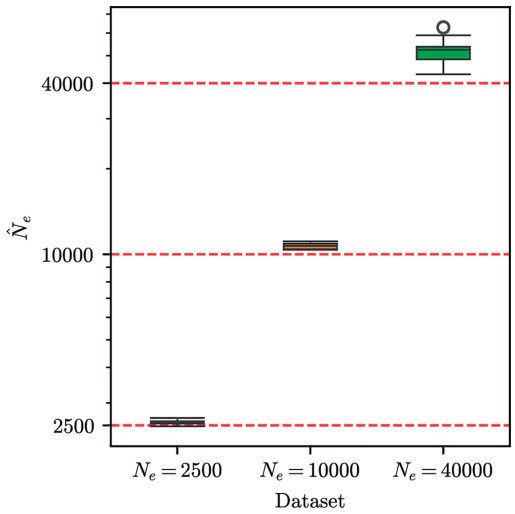

Fig 8. Estimating the effective population size N_e. Boxplots of N_e estimated using the grid-based HMM procedure, for data simulated under neutrality for $N_e \in \{2{,}500, 10{,}000, 40{,}000\}$. For each simulated value N_e, each of the 25 estimates shown are based on the composite likelihood of a batch of 10,000 replicates. N_e estimates are slightly biased upward with low variance. For all simulations, the number of generations is $T = 251$ and the initial condition is fixed at frequency $p = 0.25$. Whiskers extend to 2.5% and 97.5%-tiles.

https://doi.org/10.1371/journal.pgen.1011769.g008

Ancient DNA dataset from Great Britain

Description of GB aDNA dataset. Having demonstrated the ability of our approach to accurately characterize selection in simulated data, we applied our method to time-series genetic data obtained from human ancient DNA. To this end, we extracted the genotype information from a subset of the individuals in the Allen Ancient DNA Resource (AADR), Dataverse v7.0, which is a frequently updated repository that aims at comprising most currently published ancient DNA datasets [33].

Our method assumes that the data originates from a single panmictic population. We thus follow a rationale similar to [45] and restrict our analyses to samples from Great Britain in the last 4,450 years. We manually removed samples that were not from the mainland. Restricting to this geographic area and time window, it is unlikely that the data we analyze substantially violates the assumption of a single panmictic population, since the last major admixture event into Great Britain is estimated to have occurred before 4,450 years ago [46,47], although evidence for some more recent gene flow has been presented [48,49]. Fig R in S1 Text shows a map with sample locations and times. Moreover, Fig S in S1 Text depicts two PCA plots, demonstrating that the samples cluster with modern European individuals on a global scale, but do not exhibit evidence of strong structure on a local scale.

When analyzing samples that were genotyped using the 1240K capture assay and samples genotyped using whole genome sequencing together, we noticed spurious signals of selection, see Fig V in S1 Text. As a conservative approach, we thus only analyzed samples that were genotyped using 1240K capture, and excluded the 174 whole genome samples in this geographic area and timeframe from our dataset. We note that this conservative approach also removed the present-day samples. Our resulting dataset, henceforth referred to as the GB aDNA dataset, thus comprises 504 ancient pseudo-haploid samples genotyped using the

1240K assay and spanning 125 generations, when using a generation time of 28.1 [50]. Samples in the same generation are binned together to form the final dataset used for analysis. These individuals are a subset of the individuals published in [46,48,49].

We furthermore applied three filters to each SNP in the dataset. First, each SNP must have genotyped samples at two or more timepoints. Second, each SNP must have more than 50 (10% of 504) samples genotyped in total. Third, the minor allele frequency when pooling all samples at a given SNP across time must be greater than 0.05. We expect only SNPs that pass these filters to yield reliable signals of selection. In total, out of the 1,150,638 SNPs available, 743,417 (64.6%) pass these three filters and were used in the final analysis.

Data-matched simulations. To assess the accuracy of our method in the specific context of the GB aDNA dataset, we simulated two datasets matching the sampling scheme and timeframe of the GB aDNA dataset: In the first, which we refer to as the IBDNe dataset, we simulated the data using a history of effective population sizes that vary over time, using a population size history for the British population previously inferred from the UK10K dataset [44]. In the second, which we refer to as the const-Ne dataset, we simulate under a single constant N_e estimated from the GB aDNA dataset ($\hat{N}_e = 9{,}715$, see Significant signals of selection in the GB aDNA dataset). We show results for the IBDNe dataset in this section and show results for the dataset with constant N_e in Sect S.6 of S1 Text.

We simulated allele frequency trajectories under a particular mode of selection for $T = 125$ generations using selection coefficients $s \in \{0, 0.005, 0.01, 0.025, 0.05\}$, and sampled haplotypes given each trajectory according to the sampling times and sizes of the GB aDNA dataset. For the IBDNe dataset, we used the graphreader tool (`https://www.graphreader.com`) to extract values of N_e at each generation from Fig 4A given in [44] for the period spanned by the samples in the GB aDNA dataset. These values were then used as time-varying N_e in the Wright–Fisher simulations. Fig O of S1 Text shows the extracted N_e values. In addition, we randomly omitted sampled haplotypes with probability equal to the fraction of missingness at a randomly selected SNP in the GB aDNA dataset, to emulate the same degree of missing data. We provide a histogram of the fraction of sampled haplotypes missing for each SNP in the GB aDNA dataset in Fig T in S1 Text. To further ensure that the simulated replicates match the GB aDNA dataset, we apply the same three SNP-based filtering criteria, and only keep replicates that pass all filters. We simulated 10,000 replicates under neutrality and for each s under each of the five one-parameter selection modes.

To generate the initial frequency for each simulated replicate, we first estimated the parameters α and β for the beta distribution of the initial frequency under neutrality ($s = 0$) at each SNP in the GB aDNA dataset that passes our filters, as described in Estimation of the initial allele frequency distribution. Fig U in S1 Text shows a histogram of the mean values $\alpha/(\alpha+\beta)$ of the initial distributions estimated for each SNP. For each simulated replicate, we then chose one SNP uniformly at random, and set the initial frequency of the replicate equal to the mean of the chosen SNP. This procedure ensures that the initial frequency distribution of the simulated data matches the GB aDNA dataset closely, and captures any potential biases, for example, due to ascertainment of the 1240K SNP set.

To heuristically account for the variable population size history in the simulated data, we follow a strategy similar to [51]: We use the neutral replicates to estimate a shared constant $\hat{N}_e$ using the procedure described in Inferring effective population size, and use the inferred $\hat{N}_e$ when estimating the selection coefficients for each replicate. We then compute the MLEs for the selection coefficients using our EM-HMM for each replicate, with the mode of selection in the analysis matched to the simulated mode, and show the distribution of the MLEs $\hat{s}$ in Fig 9A. As with the simulated datasets in Simulation study, the estimates of the selection coefficients are largely unbiased. However, unlike in the simulated datasets, for strong selection,

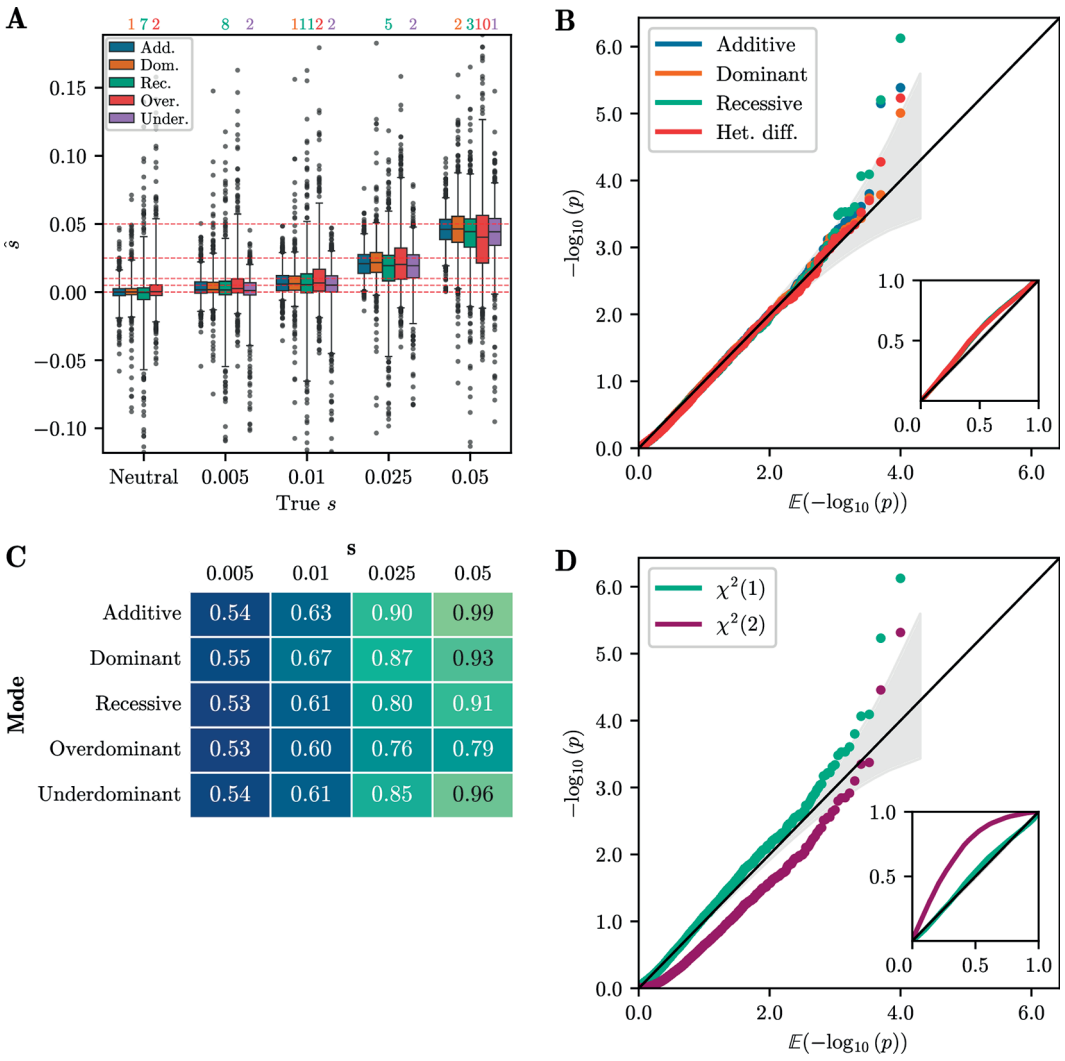

Fig 9. Accuracy of inference from data-matched simulations. A) Boxplots of MLEs $\hat{s}$ for all one-parameter selection modes. Each boxplot shows 1,000 random replicates of the 10,000 simulated. Whiskers extend to 2.5% and 97.5%-tiles. Estimates are largely unbiased for small s, and slightly biased downward for large s. B) Q-Q plot of $-\log_{10}$(p-value) against $\mathbb{E}[-\log_{10}$(p-value)] for single-alternative tests obtained using the $\chi^2(1)$ distribution. Inset shows raw p-value against expected p-value. As for the simulated datasets, p-values are well calibrated. C) Table of AUC values for data-matched simulations using likelihood-ratios for each one-parameter selection mode. D) Q-Q plot of $-\log_{10}$(p-value) against $\mathbb{E}[-\log_{10}$(p-value)] for multiple-alternative likelihood ratio statistic δ using the $\chi^2(1)$ and $\chi^2(2)$ distributions. The $\chi^2(1)$ distribution provides a good fit that is slightly anti-conservative in the tail. For all simulated replicates, the number of generations is $T = 125$, the value of N_e in each generation is derived from [44], and the initial frequencies match the GB aDNA dataset.

https://doi.org/10.1371/journal.pgen.1011769.g009

the data-matched replicates have a slight downward bias. This is likely due to the fact that the data-matched simulation comprises half the number of generations as the simulation study presented in Fig 3A. The simulated datasets with $T = 101$ generations show a similar, though less pronounced, downward bias at high s – see Fig C in S1 Text. In addition, the variance in the estimate of $\hat{s}$ for *underdominance* is lower than for the simulated datasets in Simulation study despite the shorter time horizon.

Next, we applied our procedure to test for a single alternative and our procedure to infer the mode of selection to the IBDNe simulations. For the test of a single alternative, Fig 9B shows a Q-Q plot of the p-values computed from the likelihood-ratio statistic D assuming a $\chi^2(1)$ distribution against their expected value. We again observe that these p-values are well-calibrated. As in Validation of single-alternative p-values, we computed AUC values to assess the power of the single-alternative tests on the IBDNe simulations, which are shown in Fig 9C. For $s \geq 0.025$, the AUC values are between 0.8 and 1, with the exception of *overdominant* selection. In general, the AUC values are lower than in Fig 3C; for example, for $s = 0.05$ in the case of *overdominance*, the AUC is 0.81 for the IBDNe simulations, compared to 0.99 for the simulations in Simulation study. This is likely due to the initial distribution that we used for the IBDNe-matched simulations, which has more weight close to $p = 0$, combined with the reduced number of generations. Both of these properties result in smaller cumulative changes in allele frequency, and thus power to distinguish from neutral replicates is reduced.

We also applied our multiple alternatives framework introduced in Multiple alternatives. Fig 9D shows the p-values for the neutral replicates from the IBDNe dataset, computed using the statistic δ under $\chi^2(1)$ and $\chi^2(2)$ distributions, plotted against the expected value. The p-values are in general well-calibrated when using the $\chi^2(1)$ distribution, although, as with the simulations in Validation of selection mode inference, the p-values are slightly anti-conservative in the tail of the distribution.

Fig 10 summarizes the inference of the selection mode for all replicates of the IBDNe simulations using the procedure described in Multiple alternatives; again using a p-value threshold of 0.05 to reject neutrality. For $s = 0.05$, we successfully reject neutrality for a large fraction of replicates, up to 95% of cases when simulating under *additive* selection, but the correct model is only inferred for 30% to 55% of the replicates. For the lower selection strength $s = 0.01$, neutrality is only rejected in a small percentage of the simulations. As in the simulation study presented in Validation of selection mode inference, we find that we can detect non-neutral evolution, but power to infer the correct mode of selection is limited. Note that we fail to reject 94% of neutral replicates, not 95%, due to the fact that p-values are slightly anti-conservative under the $\chi^2(1)$ distribution.

Lastly, Fig 11 shows the distribution of the inferred selection coefficient, conditional on inferring the correct mode of selection among multiple alternatives, as well as the distribution of the inferred selection coefficient for neutral replicates classified as non-neutral. For $s > 0.01$, estimated selection coefficients are similar to the unconditional estimates, indicating that our model inference procedure does not strongly bias the estimates in this parameter regime. However, for $s = 0.005$ and $s = 0.01$, most inferred coefficients are higher than the true value. As with the simulated data in Validation of selection mode inference, we observe a "winner's curse" phenomenon for lower selection coefficients, where only replicates with extreme allele frequency changes are classified as non-neutral, and consequently $\hat{s}$ is also large.

Significant signals of selection in the GB aDNA dataset. We then applied our EM-HMM algorithm and procedure to infer the mode of selection to all 743,417 SNPs in the GB aDNA dataset that pass our filters. As shown, for example, in [44], the history of effective population sizes in Great Britain varies over time. To account for this heuristically, we again mirror the strategy in [51] and first estimated a single constant effective population size $\hat{N}_e$ shared across SNPs, using the procedure described in Inferring effective population size, resulting in $\hat{N}_e = 9{,}715$. For each SNP, we then estimated the MLE using our EM-HMM for all one-parameter modes of selection, fixing the inferred $\hat{N}_e$ as the constant effective population size. As shown in Data-matched simulations, this heuristic to account for time-varying N_e yields accurate estimates of selection coefficients and well-calibrated p-values. Here, we primarily describe the results under the *additive* mode as well as the results of the procedure to

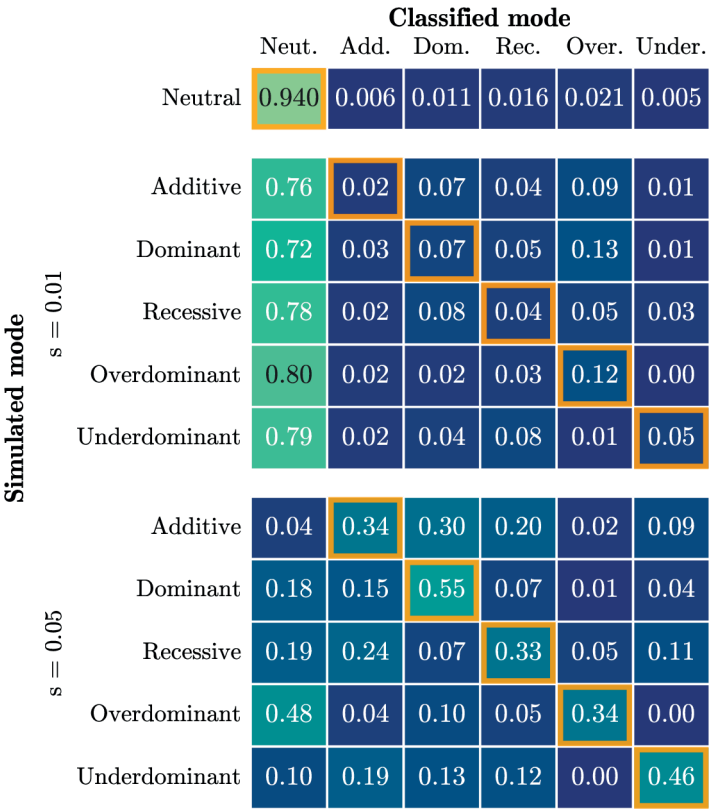

Fig 10. Confusion matrix for the procedure to infer selection mode applied to data-matched simulations. A p-value threshold of 0.05 was used. Each cell represents the fraction of replicates that were classified as a particular mode. Performance is worse than for the simulations in Simulation study – the correct model is only inferred for the plurality of replicates for $s = 0.05$. For all simulations, 10,000 replicates are simulated under a given selection mode and strength, the number of generations is $T = 125$, the value of N_e in each generation is derived from [44], and the initial frequencies match the GB aDNA dataset.

https://doi.org/10.1371/journal.pgen.1011769.g010

infer the mode of selection. Results for the other one-parameter modes, as well as Q-Q plots of the p-values for all modes, can be found in Sect S.9 of S1 Text.

Fig 12 shows a Manhattan plot of the p-values for the single-alternative likelihood-ratio test computed using the *additive* EM-HMM. We also indicate the significance threshold obtained from applying the Benjamini–Hochberg (BH) procedure [52] to correct for multiple testing at a false discovery rate (FDR) of $\alpha = 0.05$. In Sect S.10 of S1 Text, we compute the same p-values when permuting the sampling times in the dataset, demonstrating that the enrichment of low p-values we observe is a reliable signal in the data. Furthermore, we observe clusters of low p-values on Chromosomes 2, 5, and 6. This clustering of p-values is expected, since due to genetic hitchhiking [53], SNPs that are in proximity to an actual target of selection, and thus in linkage disequilibrium (LD) with the target, will also exhibit non-neutral dynamics.

In addition to these clusters, we observe several isolated SNPs with p-values exceeding the BH threshold, but no surrounding SNPs show evidence of selection, a pattern that would not be expected under genetic hitchhiking. We thus do not believe that these isolated SNPs correspond to true signals of selection and they are likely artifacts in the dataset. However, several other regions have a SNP whose p-value exceeds the BH threshold, and multiple

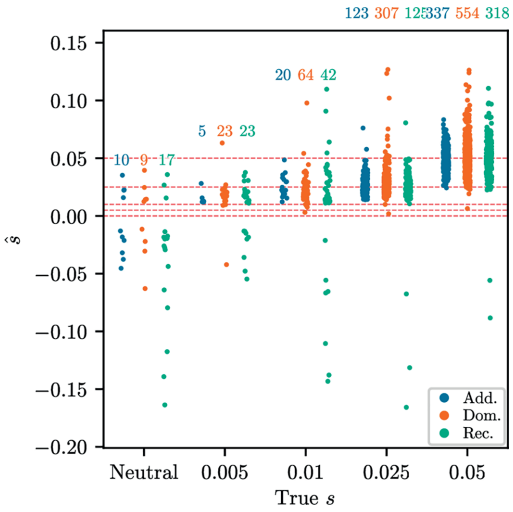

Fig 11. Strip plots of inferred $\hat{s}$ against true s for data-matched simulations. Conditioned on inferring non-neutrality for the neutral replicates and on inferring the correct selection mode among multiple alternatives for non-neutral replicates. Each strip plot is for 1,000 replicates randomly chosen from the 10,000 simulated. Number above each strip indicates the number of replicates correctly classified. For $s = 0.005$ and $s = 0.01$, the inferred values are biased upward, due to the "winner's curse". For all simulations, the number of generations is $T = 125$, the value of N_e in each generation is derived from [44], and the initial frequencies match the GB aDNA dataset.

https://doi.org/10.1371/journal.pgen.1011769.g011

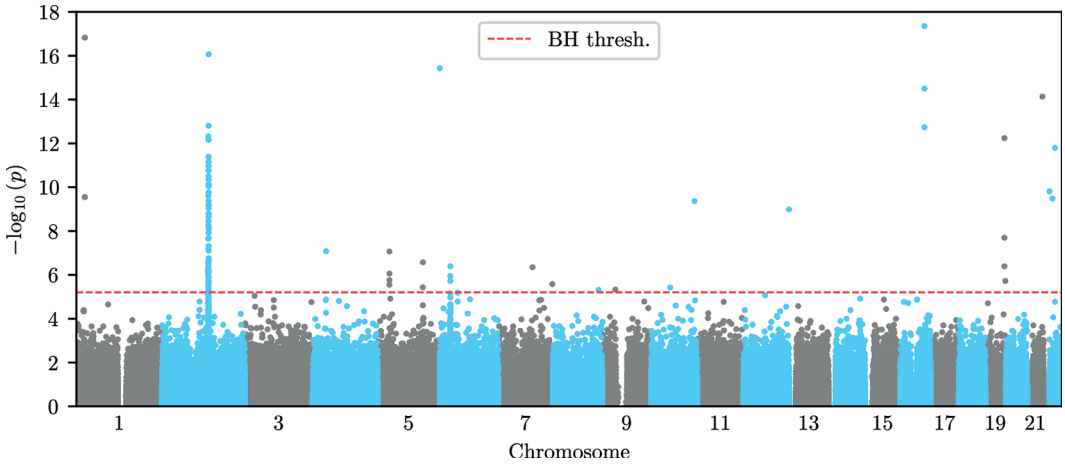

Fig 12. Manhattan plot of *additive* p-values. Manhattan plot of p-values obtained from the likelihood-ratio test under the *additive* mode of selection at all SNPs in the GB aDNA dataset. The significance threshold is obtained via the Benjamini–Hochberg procedure with an FDR of $\alpha = 0.05$. We observe clusters of significant p-values on chromosomes 2, 5, and 6, as well as several isolated signals.

https://doi.org/10.1371/journal.pgen.1011769.g012

nearby SNPs exhibiting low p-values, thus giving some support that these potentially correspond to real signals of selection. Fig 13A and 13B show the p-values in a genomic region with low p-values surrounding a significant SNP on chromosome 5 and a region surrounding a SNP with a spuriously low p-value on chromosome 7, respectively.

To remove spurious SNPs while keeping significant SNPs in regions that show additional support at surrounding SNPs, we post-process the p-values using a modified version of Brown's method [54] for combining non-independent p-values. Specifically, we consider

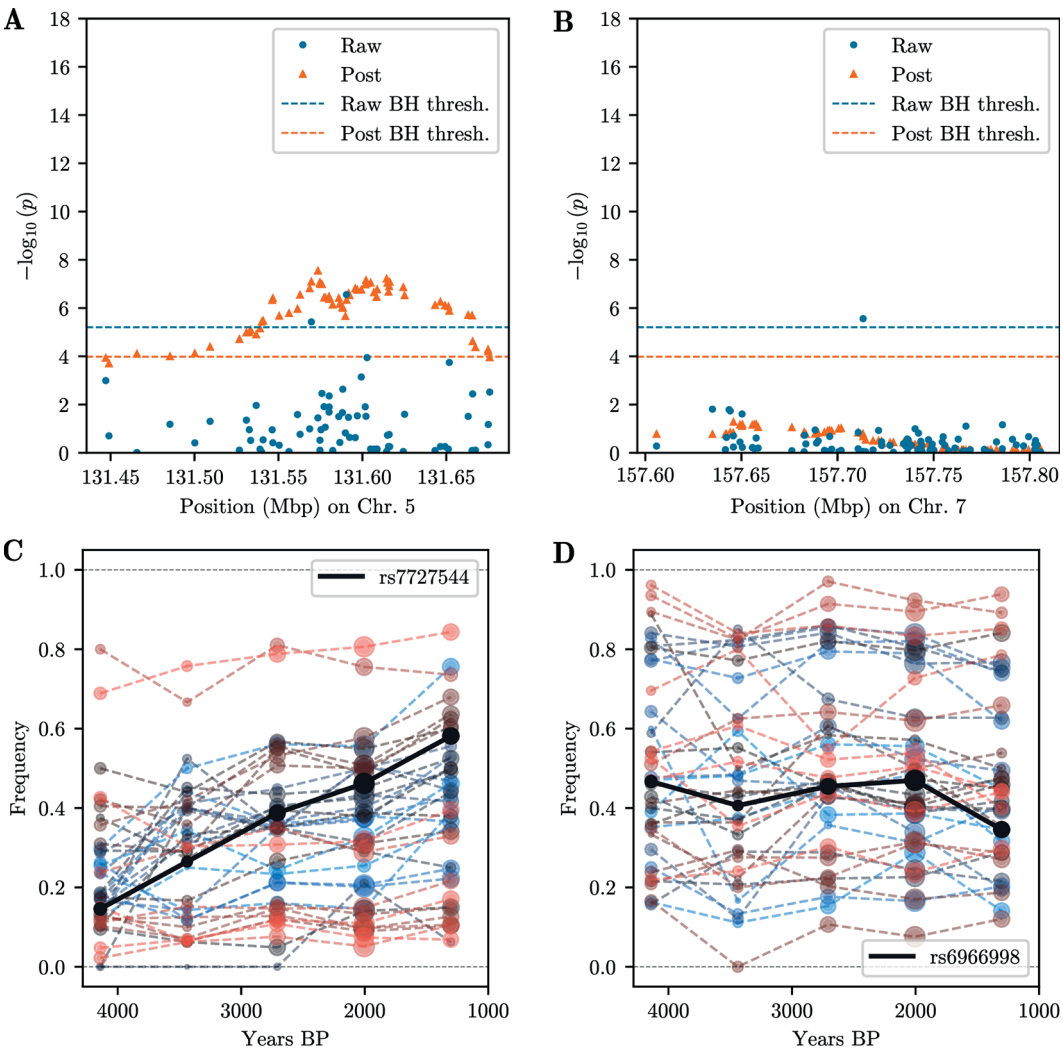

Fig 13. P-values and frequency trajectories around significant SNPs. A) Manhattan plot of p-values in genomic region on chromosome 5, centered around SNPs with p-value exceeding BH threshold. Surrounding SNPs exhibit low p-values, as expected under genetic hitchhiking. Post-processed p-values exceed the respective BH threshold. B) Manhattan plot of p-values in genomic region on chromosome 7, centered around SNP with p-value exceeding BH threshold. Surrounding SNPs do not show evidence of selection. Post-processed p-values do not show any significant signal. C) Binned allele frequency trajectories for 20 SNPs centered around SNP with lowest p-value in Fig 13A. Nearby SNPs show correlated allele frequency change, indicative of genetic hitchhiking and true signal. Size of points indicates the number of samples; color hue indicates genomic position: red smaller, blue larger. D) Binned allele frequency trajectories for 20 SNPs centered around SNP with lowest p-value in Fig 13B. Nearby SNPs do not exhibit correlated allele frequency change, suggesting spurious signal at lead SNP.

https://doi.org/10.1371/journal.pgen.1011769.g013

overlapping windows of 50 SNPs around each analyzed SNP. For each of these overlapping windows, we compute the negative sum of the logarithm of the p-values, including the focal SNP. We then fit the parameters of a scaled χ^2 distribution to these sums, and use this fitted distribution to compute post-processed p-values for each SNP. We apply the BH method to this new set of p-values to obtain a second BH threshold. Isolated SNPs, such as the SNP in

Fig 13B no longer exceed the corresponding BH threshold, while regions containing a significant SNP and additional support, such as the region in Fig 13A, have a broad peak exceeding the corresponding BH threshold.

After applying this post-processing procedure, we group significant p-values into distinct regions of significance. For a region to be considered significant, we require that at least one SNP within the region must have both a raw p-value and a post-processed p-value each exceeding its respective BH threshold. The post-processing procedure broadens p-value peaks; we therefore take each contiguous block of post-processed p-values exceeding the corresponding BH threshold as a separate candidate region. Under this criterion, for the one-dimensional *additive* EM-HMM algorithm, there are 8 distinct candidate regions. Table 1 lists these regions (in hg19 coordinates), any genes overlapping the region, the rsID of the SNP corresponding to the minimal p-value in the region, the reference and alternative alleles at this SNP, the number of significant SNPs (pre- and post-processing) in each region, the negative logarithm of the minimal p-value in the region, and the *additive* $\hat{s}$ for the derived allele at the SNP with the lowest p-value with confidence intervals. Confidence intervals were obtained by simulating 1,000 replicates matching the sampling scheme, estimated initial frequency, and estimated selection coefficient of each lead SNP, using $\hat{N}_e$, then bias-correcting all $\hat{s}$ values and taking the 0.025 and 0.975 quantiles of the simulated replicates. Additionally, we present Manhattan plots of the p-values and binned allele frequency trajectories in each significant genomic region in Sect S.11 of S1 Text.

Functional genetic variation in significant genomic regions. With the exception of TFR2, all genomic regions that we detect as significant by applying our *additive* EM-HMM have previously been identified as targets of selection, although not all have been identified as such in populations from Great Britain. In the following paragraphs, we will discuss each of these regions in the context of the relevant literature.

The strongest signal in our analysis is the well-characterized LCT locus, which has been identified as a target of selection in populations from Great Britain [45,55,56] and broader Western European populations [57–59]. The lead SNP in our analysis, rs4988235, has previously been identified as the strongest signal of selection at the LCT locus [45,58,60], and the derived allele at this SNP has been linked to the ability to digest lactase into adulthood [61]. Our estimated *additive* selection coefficient of 0.080 (CI: [0.057, 0.099]) is in line with other estimates provided in the literature [45,57,60]. Although the derived allele at this SNP has been linked to lactase persistence, recent studies argue that the introduction of milk consumption predates the increase in frequency of this allele, and that the recent strong selective pressure perhaps results from later famines where the allele proved beneficial [62,63]. Our

Table 1. Genomic regions identified as significant under *additive* selection.

Chr.	Genomic region (hg19)	Gene(s)	Lead SNP	Ref.	Alt.	Raw	Post	$-\log_{10} p_{min}$	$\hat{s}(p_{min})$
2	135,161,631-137,087,324	LCT	rs4988235	G	A	63	395	16.05	0.080 (0.057, 0.099)
2	137,164,476-137,419,024	LCT	rs580879	C	T	1	76	5.23	-0.043 (-0.061, -0.022)
4	38,680,186-38,806,462	TLR10/1/6	rs10008492	C	T	1	54	7.07	-0.047 (-0.062, -0.026)
5	33,854,740-34,004,707	SLC45A2	rs16891982	C	G	4	51	7.05	0.049 (0.029, 0.066)
5	131,465,688-131,675,046	SLC22A4	rs7727544	C	T	2	67	6.56	0.042 (0.023, 0.056)
6	31,008,368-31,091,197	HLA	rs2535317	C	T	3	151	6.38	-0.046 (-0.062, -0.025)
7	100,212,254-100,361,675	TFR2	rs4434553	A	G	1	27	6.33	0.044 (0.025, 0.061)
**16	70,771,279-71,330,074	HYDIN	rs79233902	T	G	3	51	17.34	0.242**

We indicate genes that overlap the region, number of SNPs exceeding the BH threshold pre- and post-processing, $-\log_{10} p_{min}$, and $\hat{s}$ for the derived allele at the SNP with $-\log_{10} p_{min}$. The asterisks at the HYDIN locus indicate that this signal is likely an artifact.

https://doi.org/10.1371/journal.pgen.1011769.t001

approach also identified a secondary region of significance near the LCT locus which is likely a result of genetic hitchhiking.

Genomic variation in two of our candidate regions, TLR10/1/6 and HLA, is involved in immune regulation. Polymorphisms in the TLR10/1/6 gene cluster have been linked to incidence of several cancers, as well as tuberculosis and leprosy [64–67]. The TLR10/1/6 locus has been identified as a target of selection in a previous study using ancient DNA [59], although this work pooled samples from West Eurasia. Similarly, using a dataset of present-day individuals, [68] found that the TLR genes, with the exception of TLR1, experience strong negative selection, and that the TLR10/1/6 cluster has undergone recent selection in non-African populations.

The HLA locus spans a large region on chromosome 6 and encodes a set of highly polymorphic genes critical for the function of the innate immune system. Previous studies have identified multiple signals of selection in this region [45,56], and the SNPs we identify overlap with one of these signals. Individual loci within the HLA region are believed to be under balancing or frequency-dependent selection to increase allelic diversity [69,70]. For that reason, it is somewhat surprising that genomic scans for positive selection, here and in the literature [45,56], find signals at these loci. Additionally, our scan for *heterozygote difference*, which includes *overdominance*, a model of balancing selection, reveals no significant signals in the HLA region. The likely explanation is that several alleles changing in frequency are detected as positive selection, but the short time horizon considered here is not sufficient to detect alleles under long-term balancing selection.

The SLC45A2 locus has also been well-characterized as a target of selection in European populations [71,72]. Polymorphism at this locus is associated with differences in hair and skin pigmentation [73,74]. Studies using ancient DNA data from both Western Europe broadly and Great Britain specifically have identified the SLC45A2 locus as a target of selection [45,59]. In addition, our estimated selection coefficient at the lead SNP matches values obtained from both analyses of present-day and ancient DNA: Our MLE is $s = 0.049$ (CI: [0.029, 0.066]), [72] estimate s in the range of 0.04 to 0.05, and [45] estimate $s = 0.043$.

The gene SLC22A4 is contained within the larger IBD5 locus, which consists of a group of genes with polymorphisms linked to gastrointestinal disorders such as Crohn's disease [75]. [76] found that genetic variants in SLC22A4 increases absorption of the antioxidant ergothioneine and show signals of positive selection, likely due to the low amounts of ergothioneine in early Neolithic farmer diets. Furthermore, they argue that variants linked to Crohn's disease likely increased in frequency via genetic hitchhiking. The SLC22A4/IBD5 locus was also identified as a target of selection using ancient DNA from Western Europeans [59].

Genetic variation at the TFR2 locus has not been identified as a target of selection using either contemporary or ancient genomic samples. Fig AJ in S1 Text shows that the allele frequency at the lead SNP and surrounding SNPs shift in concert indicating that this is potentially a real target of selection rather than a false signal. Mutations at the TFR2 locus cause type 3 hereditary hemochromatosis, which is characterized by abnormally high systemic iron levels [77]. In addition, a haplotype that includes the variant at the lead SNP we identify as under selection has been correlated with Parkinson's disease [78].

The HYDIN locus contains the SNP with the most significant p-value in our dataset. However, upon inspection of the p-values in this genomic region, provided in Fig AK in S1 Text, we observe that the locus indeed contains three SNPs with very low p-values, but also two 100 kbp regions without any SNPs. These empty regions are a result of our filtering procedure, which removed a large number of SNPs with minor allele frequency below 0.05. In addition, the significant SNPs at the HYDIN locus have an extremely low binned minor allele frequency

at all but the last timepoint, see Fig AK in S1 Text. Since the gene HYDIN on chromosome 16 has a pseudogene on chromosome 1 [79], these unusual patterns are potentially a result of mismapped sequence reads, and we thus believe that the signal of selection at the HYDIN locus is spurious.

Lastly, we more explicitly compare the results of our analysis to those in [45], a recent study that identified targets of selection in a temporal dataset similar to ours. The authors analyze present-day and ancient DNA samples from the AADR localized to England, dated to under 4,450 years BP, and find signals of selection in five genomic regions. Three regions (LCT, SLC45A2, HLA) are also identified in our study, three regions (TLR10/1/6, SLC22A4, TFR2) are only identified in our analysis, and two regions (DHCR7, HERC2) are only identified in [45]. Of the loci identified in both studies, the estimates of the selection coefficients at the most significant SNPs largely agree: the coefficients in the LCT, SLC45A2, and HLA regions are 0.080 (CI: [0.057, 0.099), 0.049 (CI: [0.029, 0.066]), and 0.046 (CI: [0.025, 0.061]) in our study and 0.064, 0.043, and 0.046 in [45], respectively. Figs AL and AM in S1 Text show the p-values computed using our *additive* EM-HMM and binned allele frequency trajectories in both regions identified as significant in [45] that are not significant in our analysis. At the DHCR7 locus on chromosome 11, the post-processed p-value are close to exceeding the BH threshold, but no raw p-value reaches significance. The HERC2 locus on chromosome 15 also exhibits low p-values, although to a lesser degree than DHCR7. In contrast to our conservative approach of only analyzing samples genotyped using the 1240K assay, [45] also analyze samples genotyped using whole genome data (including present-day samples), use a method that has the potential to detect selection where the *additive* coefficient changes over time, and use a different post-processing when combining signals at neighboring SNPs. Thus, a perfect alignment of signals is not expected.

Inferring the mode of selection in the GB aDNA dataset. In addition to the *additive* EM-HMM, we analyzed the GB aDNA dataset under each of the other three one-parameter modes of selection – *dominant*, *recessive*, and *heterozygote difference*, see Sect S.9 in S1 Text, where we provide full Manhattan plots for each one-parameter mode of selection, as well as tables analogous to Table 1. Additionally, we provide a table of genome-wide Spearman rank correlation coefficients between the log-likelihood ratio statistics of all one-parameter modes for the top 1% of SNPs by *additive* p-value in Fig AB of S1 Text. Both the *dominant* and the *recessive* EM-HMM have high correlation with the *additive* EM-HMM, and identify the same regions as the *additive* analysis, although the LCT locus is not split into two distinct regions in the *dominant* analysis.

In contrast, the *heterozygote difference* EM-HMM has a lower correlation with the other modes and only identifies LCT, the two SLC loci, and HYDIN as significant. The inability of this mode to detect selection at the HLA locus is at first somewhat surprising, given that balancing selection or frequency-dependent selection should result in a dynamic similar to *overdominant* selection. However, the signature that would provide strong support for these types of selection would be alleles that remain at intermediate frequencies longer than expected under neutrality, and thus the short time-horizon considered here is likely not sufficient.

We also used the procedure detailed in Multiple alternatives to infer the most likely mode of selection at each SNP. We computed the δ statistic for all SNPs in the GB aDNA dataset and used the $\chi^2(1)$ distribution to obtain p-values for each locus. We applied our procedure based on Brown's method to identify significant regions, using a BH threshold at an FDR of $\alpha = 0.05$. SNPs in the significant regions whose raw p-value exceeded the BH threshold were classified as the one-parameter mode of selection with the highest log-likelihood.

The mode inference procedure identifies the same genomic regions as the *additive* single-alternative procedure as significant. Fig 14 shows the resulting raw p-values at the LCT locus, with significant SNPs colored by their inferred selection mode. Out of the 68 SNPs exceeding the BH threshold in this region, 30 are classified as *additive*, but the other SNPs show different selection modes. The fact that a majority of loci in this region are classified as *additive* could indicate support for the hypothesis that LCT is evolving under *additive* selection; however, the inferred mode at the lead SNP rs4988235 is *dominant*. Lactase persistence functions as a dominant trait [80], lending further support to the inference of *dominant* selection at the lead SNP. Furthermore, the method presented in [45] can model non-constant selection coefficients, and the authors provide evidence that the selection at the LCT locus has weakened over time; a dynamics resembling constant *dominant* selection. We do caution against over-interpretation of these results, as the simulation study in Data-matched simulations shows that identifying even a constant mode of selection is challenging in this dataset, and thus a much greater number and density of samples is likely necessary for accurate classification. Lastly, we note that the one-parameter mode with the highest log-likelihood ratio at the lead SNP for each identified region is as follows: LCT – *dominance*, TLR10/1/6 – *additive*, SLC45A2 – *heterozygote difference*, SLC22A4 – *dominance*, HLA – *recessive*, and TFR2 – *additive*.

Coat coloration locus ASIP in domesticated horses

Description of dataset. In this section, we apply out method to a dataset presented in [34], where the authors extracted ancient DNA at six loci that affect coat coloration in domesticated horses from a set of samples in Eurasia and found evidence for selection at the ASIP and MC1R loci. Specifically, we apply our method to the ASIP locus. Fig 15A shows the sample allele frequency of the derived allele at this locus over time. The samples exhibit a

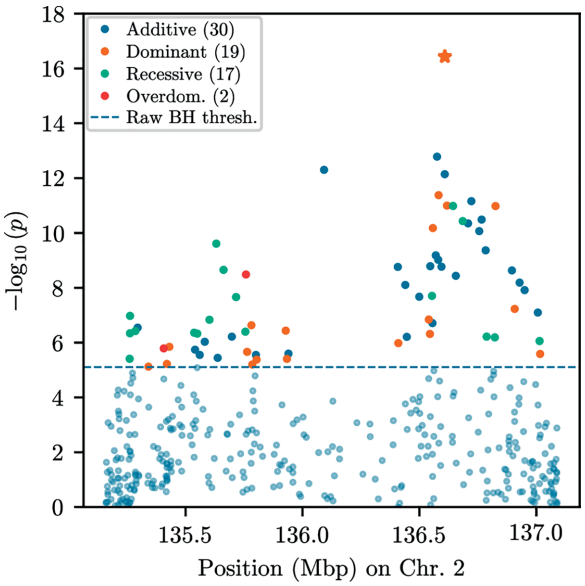

Fig 14. Manhattan plot of raw p-values at the LCT locus. P-values are computed using the procedure to identify the mode of selection described in Multiple alternatives, and significant SNPs are colored by inferred mode. The lead SNP is indicated by a larger star-shaped marker. The majority of SNPs in this region are classified as *additive*, although the lead SNP is classified as *dominant*.

https://doi.org/10.1371/journal.pgen.1011769.g014

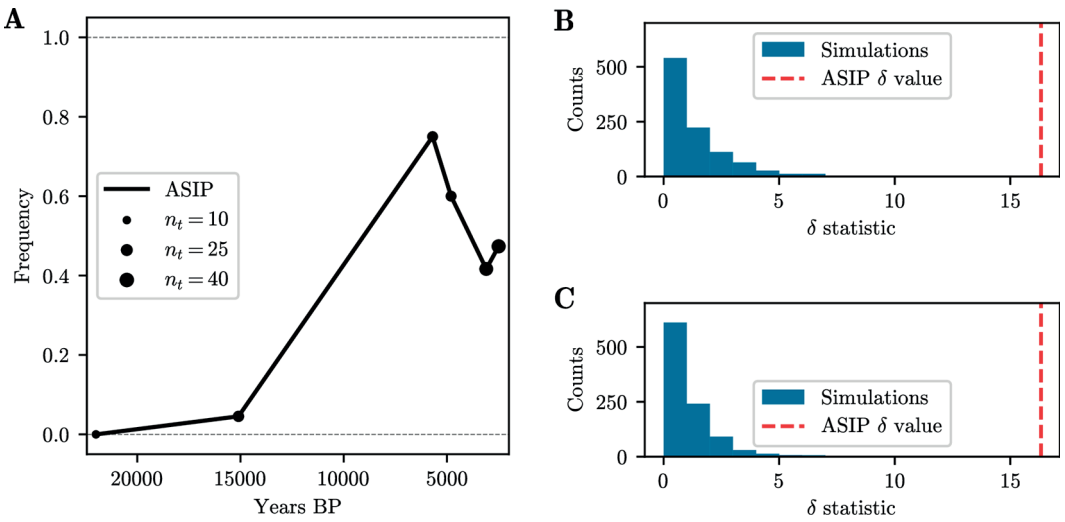

Fig 15. Frequency trajectory and evidence for non-neutrality at the ASIP locus. A) Derived allele frequency over time at the ASIP locus. The size of the points indicates the number of samples. B) & C) Histograms of δ statistic for 1,000 simulated neutral replicates matching the ASIP locus, using B) initial frequency estimated under neutrality or C) initial frequency estimated using the heterozygote difference mode. In both cases, the original dataset has a higher δ statistic (indicated by red dashed line) than any simulated replicate, providing strong evidence against neutrality.

https://doi.org/10.1371/journal.pgen.1011769.g015

sharp increase in the frequency of the derived allele, followed by leveling off at a frequency of approximately 0.5. The underlying sampled allele counts can be found in Table E of S1 Text.

This dataset has been re-analyzed in several studies under different modes of selection, with differing results; [13] analyzed the ASIP locus under *recessive* selection and did not find evidence for selection, [28] find evidence for *overdominant* selection, while more recently [81] analyze the locus under *recessive* selection and conclude that it is not under selection. The ASIP locus is known to act via a *recessive* mechanism – horses with two copies of the derived allele are black, otherwise they are bay colored [82]. It is therefore somewhat unexpected that the ASIP locus shows evidence for *overdominant* selection and not for *recessive* selection.

Single-alternative and multiple-alternative inference at the ASIP locus. We applied our EM-HMM to estimate the selection coefficients and parameters of the initial distribution at the ASIP locus under each one-parameter selection mode. Following [81] we assume N_e = 16,000 and a generation time of 8 years. For the one-parameter modes, we use the $\chi^2(1)$-based log-likelihood ratio test to compute p-values. We report the resulting estimates of the selection coefficients, the log-likelihood differences, and the p-values in Table 2. We observe that the evidence for *overdominance* selection is strongest, with a single-alternative p-value of $5.33 \cdot 10^{-5}$. In contrast to [13] and [81], we find that the ASIP locus shows evidence for *recessive* selection, although the p-value ($2.0 \cdot 10^{-2}$) is not very strong.

Although the estimated selection coefficients are low for all modes of selection (e.g. $\hat{s}$ = 0.0048 for *heterozygote difference*), the dataset comprises roughly 2,500 generations, which is over ten times as many generations as the GB aDNA dataset. This increases power to detect weaker selection; for example, the AUC values for s = 0.005 for the 100-generation simulation in Fig B of S1 Text are only slightly above 0.5, indicating minimal power to detect selection, whereas those in the 1,000-generation simulation plotted in Fig G of S1 Text are all above 0.9.

In addition to the single alternative tests, we also computed the test statistic δ for multiple alternatives (see Multiple alternatives), but used parametrized bootstrap simulations to

Table 2. Results for the ASIP locus under different selection modes.

Dataset		Add. (s_2)	Dom. (s_2)	Rec. (s_2)	Het. diff. (s_1)
Full	$\hat{s}$	0.0025	0.0022	0.0023	0.0048
	$\ell\ell - \ell\ell_0$	5.28	6.67	2.72	8.16
	p-value	$1.15 \cdot 10^{-3}$	$2.59 \cdot 10^{-4}$	$1.96 \cdot 10^{-2}$	$5.33 \cdot 10^{-5}$
Truncated	$\hat{s}$	0.0057	0.0043	0.0090	0.0047
	$\ell\ell - \ell\ell_0$	10.52	10.38	8.63	7.57
	p-value	$4.51 \cdot 10^{-6}$	$5.21 \cdot 10^{-6}$	$3.26 \cdot 10^{-5}$	$9.98 \cdot 10^{-5}$

Estimates of selection coefficients, log-likelihood differences, and p-values for all one-parameter modes. Estimates are given for both the full ASIP dataset, and the dataset truncated after the first three samples.

https://doi.org/10.1371/journal.pgen.1011769.t002

assess its significance. To this end, we simulated two sets of 1,000 neutral replicates matching the sampling scheme and number of generations to the original data at the ASIP locus. One set was simulated using the initial frequency estimated under the neutral EM-HMM and the other set with the initial frequency from the *heterozygote difference* EM-HMM, the mode with the highest likelihood among the one-parameter modes. We simulate these two sets to cover different plausible scenarios. Fig 15B and 15C show histograms of the δ statistic for the 1,000 simulated replicates using the neutral initial frequency and the *heterozygote difference* initial frequency, respectively, with the δ statistic of the original data indicated by a vertical line. In both cases, the δ statistic of the original data is much larger than any of the simulated replicates, indicating significant evidence for non-neutral dynamics. *Overdominance* selection (specifically, the *heterozygote difference* one-parameter mode with $s_1 > 0$) has the highest log-likelihood. Thus, the data at the ASIP locus supports *overdominance* selection as the most likely mode. Additionally, *recessive* selection has the lowest log-likelihood out of all one-parameter modes of selection. Since the data shows a sharp increase in frequency, followed by a plateau around 0.5, it is expected that *overdominance* is most strongly supported.

The genetic mechanism of the derived allele at the ASIP locus is *recessive*, yet the data strongly suggests that *overdominance* is the most and *recessive* selection is the least likely mode of selection. We propose the following two possible explanations for this discrepancy. First, the selection coefficient of the derived allele may have decreased during horse domestication, and some form of balancing selection may be acting after the initial increase. The point estimates in [81] indeed suggest stronger selection prior to domestication, followed by weaker selection thereafter; this is also consistent with the findings in [83]. To explore this hypothesis further, we truncated the data at the ASIP locus after the first three timepoints, to analyze just the period of frequency increase, and report the results in Table 2. We find that, indeed, the evidence for *recessive* selection is stronger than the evidence for *overdominance*, however, *additive* and *dominant* selection are also significant and have stronger p-values. A second possible explanation for our findings could be epistasis, since the derived allele at the ASIP locus has epistatic interactions with another coat coloration locus that has been shown to be under selection, the MC1R locus [82], which could affect the effective mode of selection. Future work may resolve these questions, but additional samples are likely necessary.

Discussion

In this work, we presented a novel method to compute maximum likelihood estimates (MLEs) for general diploid selection coefficients from time-series genetic data. To this end, we extended the framework in [16] for the *additive* case and derived an EM-HMM algorithm

to estimate the parameters of diploid selection. We show that the diploid EM-HMM framework can also be constrained to bespoke one-parameter models of selection via the method of Lagrange multipliers. We furthermore introduced a novel likelihood-based procedure for inferring the best fitting diploid mode of selection from temporal data between *additive*, *recessive*, *dominant*, and *over-* or *underdominant* selection. To our knowledge, our study is the first to address the statistical problem of explicitly determining the mode of selection from given time-series genetic data. Additionally, we implement a method to estimate a constant population size N_e for a given dataset, allowing for better modeling of the dynamics of genetic drift in the HMM. To further improve power to detect selection and remove spurious signals, we also introduced a procedure based on Brown's method to combine p-values across linked loci.

Using simulation studies, we show that the estimated selection coefficients are accurate across a range of selection parameters, population parameters, and sampling schemes. However, we find that determining the mode of selection from time-series data is challenging, and only yields reliable results when selection is strong. We also demonstrate that assuming *additive* selection when analyzing data simulated under different modes of selection yields comparable power to reject neutrality as when the data is simulated under *additive* selection, implying that analyzing given data assuming *additive* selection may be sufficient for scans of directional selection. However, the estimated selection coefficients are inaccurate when the mode is misspecified. In addition, we demonstrate that our procedure to account for variable population size leads to well-calibrated estimates and p-values. However, this is likely related to the short time horizon and the fact that the population size steadily increases from $N_e \approx 10{,}000$ to $N_e \approx 400{,}000$ (see Fig O in S1 Text). A more extreme history like exponential growth or severe bottlenecks will likely be more challenging, and a practitioner would have to re-assess the method in such a scenario.

We apply our method to time-series genetic data obtained from 504 ancient individuals in the AADR [33] from Great Britain dated to under 4,450 BP, and identify six genomic regions with signals of selection. These regions, except TFR2, have been identified as targets of selection in previous studies, and we discuss them in the context of the relevant literature. The regions are identified as significant under multiple directional selection modes (*additive*, *recessive*, *dominant*). When classifying the mode of selection from the data, however, the results are inconclusive: For example, we find that a majority of the SNPs at the LCT locus provide evidence for *additive* selection, but the lead SNPs is classified as *dominant*. In addition, we reanalyze a time-series dataset consisting of 146 samples over 2,400 generations from the ASIP locus involved in coat coloration in horses, and show evidence for selection under different non-additive modes.

Note that our HMM-implementation uses the Chebychev nodes to compute the single-generation transition matrix accurately across the hidden state space, and the integrals account for the probability mass that is absorbed in the boundary. Capturing these features accurately is important when implementing the model [17,84], and consequently, we believe that our algorithm identifies the MLEs from a given temporal datasets and computes the likelihoods with high accuracy. Thus, the statistical properties of the MLEs and the likelihood-ratio tests exhibited in our simulation studies in Simulation study and Data-matched simulations are likely not exclusive to our method but potentially characterize the MLEs and the power of the respective tests in general under the given population genetic model, regardless of whether our method or a different likelihood-based method is used for the analysis.

Moreover, we characterize the statistical properties in a range of scenarios, but if these scenarios do not cover the exact scheme encountered in a specific empirical dataset, our simulation framework can be readily modified to characterize the statistical properties in the

respective scenario. Naturally, the power to identify and characterize selection does depend on the exact sampling scheme: Strong selection can be readily identified, even when samples are limited to a short time period. However, weaker selection requires sampling more data over a longer time period. For example, in our analysis of the GB aDNA dataset, Figs 9C and 10 demonstrate limited power to detect selection as strong as $s = 0.01$ in the respective scenario.

Our method computes the MLEs of general diploid selection parameters, and we believe that this is useful to researchers in at least two regards: (1) Our approach can be used to infer the mode of selection from a given temporal dataset. While we demonstrate that selection needs to be fairly strong for reliable classification, our framework can be used to characterize statistical power in a given scenario, and determine whether additional samples at potentially additional timepoints are necessary. (2) If the selection mode operating on the genetic variants in a given temporal dataset is known a priori, for example, *dominance* at the LCT locus or *underdominance* dynamics resulting from stabilizing selection on complex traits, our method enables researchers to estimate the selection coefficients accurately under the correct model. We demonstrate that assuming the wrong mode of selection can yield inaccurate estimates.

In practice, we recommend the following approach to analyze given time-series genetic data, potentially at a large number of loci: If computational resources are limited, researchers should apply the *additive* EM-HMM to obtain MLEs of the *additive* selection coefficient at each locus, and use standard likelihood-ratio testing to identify outliers. As detailed in Robustness of selection coefficient estimation, the likelihood-ratio test under *additive* selection can identify non-neutral replicates, even if the mode of selection is misspecified, but estimated coefficients are inaccurate. For reference, the *additive* analysis of the 743,417 SNPs for 504 samples over 125 generations in our GB aDNA dataset took roughly 5,000 cpu-hours. If additional computational resources are available, we recommend analyzing the data under each bespoke one-parameter selection mode, as well as the unconstrained mode, to characterize signals in the data that are not correctly described by *additive* selection. The results can then also be used to identify the mode of selection from the data, but as we demonstrate, accuracy is limited. In addition, we strongly recommend performing a data-matched simulation study, as presented in Data-matched simulations or Sect S.6 in S1 Text. Such data-matched simulations enable exact characterization of the statistical power and accuracy of the approach in the specific scenario.

As for similar approaches, the HMM underlying our method assumes that the population is panmictic, and violations of this assumption can dilute signals or introduce spurious signals. Future work to address this shortcoming could proceed along at least two possible avenues: (1) Controlling for population structure by using Principal Components as covariates in the estimation procedure directly [85,86], or (2) explicitly including the population structure and exchange of migrants in the underlying population genetic model [87].

Moreover, our approach estimates selection coefficients from temporal data at the focal locus only, and does not incorporate the allele frequency dynamics at linked loci. In our analysis of the GB aDNA dataset, we do leverage signals across loci using a novel post-processing approach to combine p-values in a genomic window. This post-processing can reduce signal-to-noise ratios in genomic scans for selection in general. However, incorporating the genetic variation at multiple SNPs using a proper likelihood model for the multi-locus dynamics under the Wright-Fisher process has the potential to account for chromosomal linkage more accurately and result in a more robust inference [88,89].

While we focus on analyzing time-series genetic data at a single locus in this study, the capacity of our method to characterize selection modes more general than *additive* only also

has potential benefits when studying polygenic selection on complex traits: In models of stabilizing polygenic selection around an optimal trait, the genetic variants affecting the trait experience *underdominant* selection dynamics, which can be readily addressed using our framework.

Supporting information

S1 Text. Supplementary Material. Document containing additional details of the method, as well as figures and tables to supplement the analyses.
(PDF)

Acknowledgments

We want to thank Xinyi Li, Xiaoheng Cheng, Constanza de la Fuente, and Maanasa Raghavan for helpful comments on the method and the data analysis. Moreover, we thank Jeremy Berg, Maryn Carlson and Rowan Hart for comments on the manuscript. In addition, we thank the members of the Raghavan, Berg, and Novembre labs for valuable feedback throughout the project.

Author contributions

Conceptualization: Matthias Steinrücken.

Data curation: Adam G. Fine, Matthias Steinrücken.

Formal analysis: Adam G. Fine, Matthias Steinrücken.

Funding acquisition: Matthias Steinrücken.

Investigation: Adam G. Fine, Matthias Steinrücken.

Methodology: Adam G. Fine, Matthias Steinrücken.

Project administration: Adam G. Fine, Matthias Steinrücken.

Resources: Matthias Steinrücken.

Software: Adam G. Fine, Matthias Steinrücken.

Supervision: Matthias Steinrücken.

Validation: Adam G. Fine, Matthias Steinrücken.

Visualization: Adam G. Fine, Matthias Steinrücken.

Writing – original draft: Adam G. Fine.

Writing – review & editing: Adam G. Fine, Matthias Steinrücken.

References

1. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437(7062):1153–7. https://doi.org/10.1038/nature04240 PMID: 16237444
2. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, et al. Signatures of mutation and selection in the cancer genome. Nature. 2010;463(7283):893–8. https://doi.org/10.1038/nature08768 PMID: 20164919
3. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005;39:197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420 PMID: 16285858

4. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97–120. https://doi.org/10.1146/annurev-genet-111212-133526 PMID: 24274750
5. Lachance J, Tishkoff SA. Population genomics of human adaptation. Annu Rev Ecol Evol Syst. 2013;44:123–43. https://doi.org/10.1146/annurev-ecolsys-110512-135833 PMID: 25383060
6. Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockhammer PW Avila-Arcos MC, et al. Ancient DNA analysis. Nat Rev Methods Primers. 2021;1(1):1–26. https://doi.org/10.1038/s43586-021-00016-3
7. Hofreiter M, Paijmans JLA, Goodchild H, Speller CF, Barlow A, Fortes GG, et al. The future of ancient DNA: technical advances and conceptual shifts. BioEssays. 2015;37(3):284–93. https://doi.org/10.1002/bies.201400160 PMID: 25413709
8. Barghi N, Tobler R, Nolte V, Jakšić AM, Mallard F, Otte KA, et al. Genetic redundancy fuels polygenic adaptation in drosophila. PLOS Biol. 2019;17(2):e3000128. https://doi.org/10.1371/journal.pbio.3000128 PMID: 30716062
9. Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity (Edinb). 2015;114(5):431–40. https://doi.org/10.1038/hdy.2014.86 PMID: 25269380
10. Malaspinas A-S. Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. Mol Ecol. 2016;25(1):24–41. https://doi.org/10.1111/mec.13492 PMID: 26613371
11. Dehasque M, Ávila-Arcos MC, Díez-Del-Molino D, Fumagalli M, Guschanski K, Lorenzen ED, et al. Inference of natural selection from ancient DNA. Evol Lett. 2020;4(2):94–108. https://doi.org/10.1002/evl3.165 PMID: 32313686
12. Bollback JP, York TL, Nielsen R. Estimation of 2Nes from temporal allele frequency data. Genetics. 2008;179(1):497–502. https://doi.org/10.1534/genetics.107.085019 PMID: 18493066
13. Malaspinas A-S, Malaspinas O, Evans SN, Slatkin M. Estimating allele age and selection coefficient from time-serial data. Genetics. 2012;192(2):599–607. https://doi.org/10.1534/genetics.112.140939 PMID: 22851647
14. Wang J. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. Genet Res. 2001;78(3):243–57. https://doi.org/10.1017/s0016672301005286 PMID: 11865714
15. Ferrer-Admetlla A, Leuenberger C, Jensen JD, Wegmann D. An approximate markov model for the wright-fisher diffusion and its application to time series data. Genetics. 2016;203(2):831–46. https://doi.org/10.1534/genetics.115.184598 PMID: 27038112
16. Mathieson I, McVean G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics. 2013;193(3):973–84. https://doi.org/10.1534/genetics.112.147611 PMID: 23307902
17. Tataru P, Simonsen M, Bataillon T, Hobolth A. Statistical inference in the Wright-Fisher model using allele frequency data. Syst Biol. 2017;66(1):e30–46. https://doi.org/10.1093/sysbio/syw056 PMID: 28173553
18. Vlachos C, Burny C, Pelizzola M, Borges R, Futschik A, Kofler R, et al. Benchmarking software tools for detecting and quantifying selection in evolve and resequencing studies. Genome Biol. 2019;20(1):169. https://doi.org/10.1186/s13059-019-1770-8 PMID: 31416462
19. Gemmell NJ, Slate J. Heterozygote advantage for fecundity. PLoS One. 2006;1(1):e125. https://doi.org/10.1371/journal.pone.0000125 PMID: 17205129
20. Hedrick PW. What is the evidence for heterozygote advantage selection?. Trends Ecol Evol. 2012;27(12):698–704. https://doi.org/10.1016/j.tree.2012.08.012 PMID: 22975220
21. Palmer DS, Zhou W, Abbott L, Wigdor EM, Baya N, Churchhouse C, et al. Analysis of genetic dominance in the UK Biobank. Science. 2023;379(6639):1341–8. https://doi.org/10.1126/science.abn8455 PMID: 36996212
22. Sanjak JS, Sidorenko J, Robinson MR, Thornton KR, Visscher PM. Evidence of directional and stabilizing selection in contemporary humans. Proc Natl Acad Sci U S A. 2018;115(1):151–6. https://doi.org/10.1073/pnas.1707227114 PMID: 29255044
23. Barton N. The maintenance of polygenic variation through a balance between mutation and stabilizing selection. Genet Res. 1986;47(3):209–16. https://doi.org/10.1017/s0016672300023156 PMID: 3744046
24. de Vladar HP, Barton N. Stability and response of polygenic traits to stabilizing selection and mutation. Genetics. 2014;197(2):749–67. https://doi.org/10.1534/genetics.113.159111 PMID: 24709633
25. Simons YB, Bullaughey K, Hudson RR, Sella G. A population genetic interpretation of GWAS findings for human quantitative traits. PLoS Biol. 2018;16(3):e2002985. https://doi.org/10.1371/journal.pbio.2002985 PMID: 29547617

26. Koch E, Connally NJ, Baya N, Reeve MP, Daly M, Neale B, et al. Genetic association data are broadly consistent with stabilizing selection shaping human common diseases and traits. bioRxiv. 2024.06.19.599789. https://doi.org/10.1101/2024.06.19.599789
27. Cheng X, Steinrücken M. diplo-locus: A lightweight toolkit for inference and simulation of time-series genetic data under general diploid selection. bioRxiv. 2023.10.12.562101. https://doi.org/10.1101/2023.10.12.562101 PMID: 37905072
28. Steinrücken M, Bhaskar A, Song YS. A novel spectral method for inferring general diploid selection from time series genetic data. Ann Appl Stat. 2014;8(4):2203–22. https://doi.org/10.1214/14-aoas764 PMID: 25598858
29. Foll M, Shim H, Jensen JD. WFABC: a Wright–Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Mol Ecol Resour. 2015;15(1):87–98. https://doi.org/10.1111/1755-0998.12280 PMID: 24834845
30. Schraiber JG, Evans SN, Slatkin M. Bayesian inference of natural selection from allele frequency time series. Genetics. 2016;203(1):493–511. https://doi.org/10.1534/genetics.116.187278 PMID: 27010022
31. Iranmehr A, Akbari A, Schlötterer C, Bafna V. Clear: composition of likelihoods for evolve and resequence experiments. Genetics. 2017;206(2):1011–23. https://doi.org/10.1534/genetics.116.197566 PMID: 28396506
32. Taus T, Futschik A, Schlötterer C. Quantifying selection with pool-seq time series data. Mol Biol Evol. 2017;34(11):3023–34. https://doi.org/10.1093/molbev/msx225 PMID: 28961717
33. Mallick S, Micco A, Mah M, Ringbauer H, Lazaridis I, Olalde I, et al. The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. Sci Data. 2024;11(1):182. https://doi.org/10.1038/s41597-024-03031-7 PMID: 38341426
34. Ludwig A, Pruvost M, Reissmann M, Benecke N, Brockmann GA, Castaños P, et al. Coat color variation at the beginning of horse domestication. Science. 2009;324(5926):485. https://doi.org/10.1126/science.1172750 PMID: 19390039
35. Ewens WJ. Mathematical population genetics. 2 ed. Springer; 2004.
36. Watterson GA. Testing selection at a single locus. Biometrics. 1982;38(2):323–31. https://doi.org/10.2307/2530446 PMID: 7115865
37. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989;77(2):257–86. https://doi.org/10.1109/5.18626
38. Bishop CM. Pattern recognition and machine learning. Springer; 2006.
39. Hoffmann LD, Bradley GL, Rosen KH. Applied calculus for business, economics, and the social and life sciences. Expanded 10th ed. McGraw-Hill; 2010.
40. Mathews J, Fink K. Numerical methods using matlab. 4th ed. Pearson; 2003.
41. Varadhan R, Roland C. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. Scand J Stat. 2008;35(2):335–53. https://doi.org/10.1111/j.1467-9469.2007.00585.x
42. Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Statist. 1938;9(1):60–2. https://doi.org/10.1214/aoms/1177732360
43. Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. Genetics. 1992;132(4):1161–76. https://doi.org/10.1093/genetics/132.4.1161 PMID: 1459433
44. Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am J Hum Genet. 2015;97(3):404–18. https://doi.org/10.1016/j.ajhg.2015.07.012 PMID: 26299365
45. Mathieson I, Terhorst J. Direct detection of natural selection in Bronze Age Britain. Genome Res. 2022;32(11–12):2057–67. https://doi.org/10.1101/gr.276862.122 PMID: 36316157
46. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, et al. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature. 2018;555(7695):190–6. https://doi.org/10.1038/nature25738 PMID: 29466337
47. Chintalapati M, Patterson N, Moorjani P. The spatiotemporal patterns of major human admixture events during the European Holocene. eLife. 2022;11:e77625. https://doi.org/10.7554/elife.77625 PMID: 35635751
48. Patterson N, Isakov M, Booth T, Büster L, Fischer C-E, Olalde I, et al. Large-scale migration into britain during the middle to late bronze age. Nature. 2022;601(7894):588–94. https://doi.org/10.1038/s41586-021-04287-4 PMID: 34937049
49. Gretzinger J, Sayer D, Justeau P, Altena E, Pala M, Dulias K, et al. The Anglo-Saxon migration and the formation of the early English gene pool. Nature. 2022;610(7930):112–9. https://doi.org/10.1038/s41586-022-05247-2 PMID: 36131019

50. Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. Proc Natl Acad Sci U S A. 2016;113(20):5652–7. https://doi.org/10.1073/pnas.1514696113 PMID: 27140627
51. Foll M, Poh Y-P, Renzette N, Ferrer-Admetlla A, Bank C, Shim H, et al. Influenza virus drug resistance: a time-sampled population genetics perspective. PLoS Genet. 2014;10(2):e1004185. https://doi.org/10.1371/journal.pgen.1004185 PMID: 24586206
52. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B: Statist Methodol. 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
53. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23(1):23–35. https://doi.org/10.1017/S0016672300014634 PMID: 4407212
54. Brown MB. A method for combining non-independent, one-sided tests of significance. Biometrics. 1975;31(4):987–92. https://doi.org/10.2307/2529826
55. Nait Saada J, Kalantzis G, Shyr D, Cooper F, Robinson M, Gusev A, et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nat Commun. 2020;11(1):6130. https://doi.org/10.1038/s41467-020-19588-x PMID: 33257650
56. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. Science. 2016;354(6313):760–4. https://doi.org/10.1126/science.aag0776 PMID: 27738015
57. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004;74(6):1111–20. https://doi.org/10.1086/421051 PMID: 15114531
58. Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG. The origins of lactase persistence in Europe. PLoS Comput Biol. 2009;5(8):e1000491. https://doi.org/10.1371/journal.pcbi.1000491 PMID: 19714206
59. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528(7583):499–503. https://doi.org/10.1038/nature16152 PMID: 26595274
60. Peter BM, Huerta-Sanchez E, Nielsen R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. 2012;8(10):e1003011. https://doi.org/10.1371/journal.pgen.1003011 PMID: 23071458
61. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. Identification of a variant associated with adult-type hypolactasia. Nat Genet. 2002;30(2):233–7. https://doi.org/10.1038/ng826 PMID: 11788828
62. Burger J, Link V, Blocher J, Schulz A, Sell C, Pochon Z, et al. Low prevalence of lactase persistence in bronze age Europe indicates ongoing strong selection over the last 3,000 years. Curr Biol. 2020;30(21):4307–15. https://doi.org/10.1016/j.cub.2020.08.033 PMID: 32888485
63. Evershed RP, Davey Smith G, Roffet-Salque M, Timpson A, Diekmann Y, Lyon MS, et al. Dairying, diseases and the evolution of lactase persistence in Europe. Nature. 2022;608:336–45. https://doi.org/10.1038/s41586-022-05010-7 PMID: 35896751
64. Purdue MP, Lan Q, Wang SS, Kricker A, Menashe I, Zheng T-Z, et al. A pooled investigation of Toll-like receptor gene variants and risk of non-Hodgkin lymphoma. Carcinogenesis. 2009;30(2):275–81. https://doi.org/10.1093/carcin/bgn262 PMID: 19029192
65. Sun J, Wiklund F, Hsu F-C, Bälter K, Zheng SL, Johansson J-E, et al. Interactions of sequence variants in interleukin-1 receptor-associated kinase4 and the toll-like receptor 6-1-10 gene cluster increase prostate cancer risk. Cancer Epidemiol Biomarkers Prev. 2006;15(3):480–5. https://doi.org/10.1158/1055-9965.EPI-05-0645 PMID: 16537705
66. Ma X, Liu Y, Gowen BB, Graviss EA, Clark AG, Musser JM. Full-exon resequencing reveals toll-like receptor variants contribute to human susceptibility to tuberculosis disease. PLoS One. 2007;2(12):e1318. https://doi.org/10.1371/journal.pone.0001318 PMID: 18091991
67. Wong SH, Gochhait S, Malhotra D, Pettersson FH, Teo YY, Khor CC, et al. Leprosy and the adaptation of human toll-like receptor 1. PLoS Pathog. 2010;6(7):e1000979. https://doi.org/10.1371/journal.ppat.1000979 PMID: 20617178
68. Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, et al. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. 2009;5(7):e1000562. https://doi.org/10.1371/journal.pgen.1000562 PMID: 19609346
69. Hedrick PW, Thomson G. Evidence for balancing selection at HLA. Genetics. 1983;104(3):449–56. https://doi.org/10.1093/genetics/104.3.449 PMID: 6884768

70. Bronson PG, Mack SJ, Erlich HA, Slatkin M. A sequence-based approach demonstrates that balancing selection in classical human leukocyte antigen (HLA) loci is asymmetric. Hum Mol Genet. 2013;22(2):252–61. https://doi.org/10.1093/hmg/dds424 PMID: 23065702
71. Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet. 2007;71:354–69. https://doi.org/10.1111/j.1469-1809.2006.00341.x PMID: 17233754
72. Beleza S, Santos AM, McEvoy B, Alves I, Martinho C, Cameron E, et al. The timing of pigmentation lightening in Europeans. Mol Biol Evol. 2013;30(1):24–35. https://doi.org/10.1093/molbev/mss207 PMID: 22923467
73. Soejima M, Koda Y. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. Int J Legal Med. 2007;121(1):36–9. https://doi.org/10.1007/s00414-006-0112-z PMID: 16847698
74. Hysi PG, Valdes AM, Liu F, Furlotte NA, Evans DM, Bataille V, et al. Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. Nat Genet. 2018;50(5):652–6. https://doi.org/10.1038/s41588-018-0100-5 PMID: 29662168
75. Fisher SA, Hampe J, Onnie CM, Daly MJ, Curley C, Purcell S, et al. Direct or indirect association in a complex disease: the role of SLC22A4 and SLC22A5 functional variants in Crohn disease. Hum Mutat. 2006;27(8):778–85. https://doi.org/10.1002/humu.20358 PMID: 16835882
76. Huff CD, Witherspoon DJ, Zhang Y, Gatenbee C, Denson LA, Kugathasan S, et al. Crohn's disease and genetic hitchhiking at IBD5. Mol Biol Evol. 2012;29(1):101–11. https://doi.org/10.1093/molbev/msr151 PMID: 21816865
77. Girelli D, Bozzini C, Roetto A, Alberti F, Daraio F, Colombari R, et al. Clinical and pathologic findings in hemochromatosis type 3 due to a novel mutation in transferrin receptor 2 gene. Gastroenterology. 2002;122(5):1295–302. https://doi.org/10.1053/gast.2002.32984 PMID: 11984516
78. Rhodes SL, Buchanan DD, Ahmed I, Taylor KD, Loriot M-A, Sinsheimer JS, et al. Pooled analysis of iron-related genes in Parkinson's disease: association with transferrin. Neurobiol Dis. 2014;62:172–8. https://doi.org/10.1016/j.nbd.2013.09.019 PMID: 24121126
79. Dutcher SK, Brody SL. HY-DIN' in the Cilia: discovery of central pair-related mutations in primary ciliary dyskinesia. Am J Respir Cell Mol Biol. 2020;62(3):281–2. https://doi.org/10.1165/rcmb.2019-0316ED PMID: 31604022
80. Swallow DM. Genetics of lactase persistence and lactose intolerance. Annu Rev Genet. 2003;37:197–219. https://doi.org/10.1146/annurev.genet.37.110801.143820 PMID: 14616060
81. He Z, Dai X, Lyu W, Beaumont M, Yu F. Estimating temporally variable selection intensity from ancient DNA data. Mol Biol Evol. 2023;40(3):msad008. https://doi.org/10.1093/molbev/msad008 PMID: 36661852
82. Rieder S, Taourit S, Mariat D, Langlois B, Guérin G. Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus). Mamm Genome. 2001;12(6):450–5. https://doi.org/10.1007/s003350020017 PMID: 11353392
83. Wutke S, Benecke N, Sandoval-Castellanos E, Döhle H-J, Friederich S, Gonzalez J, et al. Spotted phenotypes in horses lost attractiveness in the Middle Ages. Sci Rep. 2016;6:38548. https://doi.org/10.1038/srep38548 PMID: 27924839
84. Paris C, Servin B, Boitard S. Inference of selection from genetic time series using various parametric approximations to the Wright-Fisher Model. G3 (Bethesda). 2019;9(12):4073–86. https://doi.org/10.1534/g3.119.400778 PMID: 31597676
85. Luu K, Bazin E, Blum MGB. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour. 2017;17(1):67–77. https://doi.org/10.1111/1755-0998.12592 PMID: 27601374
86. Ju D, Mathieson I. The evolution of skin pigmentation-associated variation in West Eurasia. Proc Natl Acad Sci U S A. 2021;118(1):e2009227118. https://doi.org/10.1073/pnas.2009227118 PMID: 33443182
87. Joseph TA, Pe'er I. Inference of population structure from time-series genotype data. Am J Hum Genet. 2019;105(2):317–33. https://doi.org/10.1016/j.ajhg.2019.06.002 PMID: 31256878
88. Terhorst J, Schlötterer C, Song YS. Multi-locus analysis of genomic time series data from experimental evolution. PLoS Genet. 2015;11(4):e1005069. https://doi.org/10.1371/journal.pgen.1005069 PMID: 25849855
89. He Z, Dai X, Beaumont M, Yu F. Detecting and quantifying natural selection at two linked loci from time series data of allele frequencies with Forward-in-time simulations. Genetics. 2020;216:521–41. https://doi.org/10.1534/genetics.120.303463 PMID: 32826299