Fast and Accurate Estimation of Selection Coefficients and Allele Histories from Ancient and Modern DNA

Andrew H. Vaughn [1,*] Rasmus Nielsen [2,3]

[1]Center for Computational Biology, University of California, Berkeley, CA 94720, USA
[2]Departments of Integrative Biology and Statistics, University of California, Berkeley, CA 94720, USA
[3]Center for GeoGenetics, University of Copenhagen, Copenhagen DK-1350, Denmark

***Corresponding author**: E-mail: ahv36@berkeley.edu.

Associate editor: Yoseob Kim

Abstract

We here present CLUES2, a full-likelihood method to infer natural selection from sequence data that is an extension of the method CLUES. We make several substantial improvements to the CLUES method that greatly increases both its applicability and its speed. We add the ability to use ancestral recombination graphs on ancient data as emissions to the underlying hidden Markov model, which enables CLUES2 to use both temporal and linkage information to make estimates of selection coefficients. We also fully implement the ability to estimate distinct selection coefficients in different epochs, which allows for the analysis of changes in selective pressures through time, as well as selection with dominance. In addition, we greatly increase the computational efficiency of CLUES2 over CLUES using several approximations to the forward–backward algorithms and develop a new way to reconstruct historic allele frequencies by integrating over the uncertainty in the estimation of the selection coefficients. We illustrate the accuracy of CLUES2 through extensive simulations and validate the importance sampling framework for integrating over the uncertainty in the inference of gene trees. We also show that CLUES2 is well-calibrated by showing that under the null hypothesis, the distribution of log-likelihood ratios follows a χ^2 distribution with the appropriate degrees of freedom. We run CLUES2 on a set of recently published ancient human data from Western Eurasia and test for evidence of changing selection coefficients through time. We find significant evidence of changing selective pressures in several genes correlated with the introduction of agriculture to Europe and the ensuing dietary and demographic shifts of that time. In particular, our analysis supports previous hypotheses of strong selection on lactase persistence during periods of ancient famines and attenuated selection in more modern periods.

Key words: **ARGs, ancient DNA, selection, HMMs, lactase persistence.**

Methods

Introduction

One of the primary evolutionary forces that shapes the genetic variation of populations is natural selection, the causal effect of genotype on the reproductive success of an individual. While experimental evolution studies can enable a somewhat direct measurement of natural selection in certain limited cases, most methods for inferring natural selection rely on statistical frameworks to analyze a sample of sequence data. Given that sequence data lies in a very high-dimensional space, these methods for inferring natural selection have historically been based on summary statistics, that can be thought of as projections of the sequence data to a lower-dimensional subspace. These summary statistics include single nucleotide polymorphism (SNP)-frequency statistics such as Tajima's D (Tajima 1989) or Fay and Wu's H (Fay and Wu 2000), as well as linkage disequilibrium statistics such as the extended haplotype homozygosity statistic (Sabeti et al. 2002) or the LD decay test (Wang et al. 2006). One can compute these statistics for a given sequence, or in sliding windows across the genome, and conclude that values significantly different from their null expectation are indicative of natural selection. As an extension of the basic summary statistic approach, approximate Bayesian computation (ABC) frameworks, which can use information from many different summary statistics, have also been used to infer selection coefficients (Peter et al. 2012). A variety of machine learning methods to detect selection have additionally been developed (Schrider and Kern 2018; Torada et al. 2019; Hejase et al. 2021), which all rely on training models on a set of features obtained from sequence data. However, these approaches ignore a large amount of the information present in the data by only considering sets of summary statistics or features. Furthermore, these methods lack flexibility, for example if one wishes to specify different parameters or classes of models, such as allowing for selection coefficients to differ in different time periods or incorporating ancient samples that are sampled at distinct time points.

Received: December 16, 2023. **Revised:** July 02, 2024. **Accepted:** July 10, 2024

Overall, therefore, a computationally tractable full-likelihood method is needed that does not rely on summary statistics, and can thus use all the information present in the data, and which also has the flexibility to be able to be applied to a variety of datasets and models. It is to this end that the method CLUES was developed by Stern et al. (2019). The original CLUES method is based on the observation that if one can determine a full-likelihood expression for the data conditional on the full history of the frequency of an allele, then one can consider the allele history as a latent variable and integrate over the full set of possible allele frequencies in order to obtain an expression for the full-likelihood function. As the sequence data and selection coefficient of an allele are approximately independent conditional on the gene tree at that locus, the full likelihood of the data could be computed by computing the likelihood of that gene tree. In reality, the true gene tree at a site is never known with certainty but instead must be estimated. Estimating gene trees in recombining species is challenging as recombination causes neighboring regions of the genome to have distinct, but correlated, gene trees. The full collection of these correlated gene trees across a genomic region is called the ancestral recombination graph (ARG), the theory for which was first developed by Hudson (1983) and Griffiths and Marjoram (1996). The key insight of using ARGs for population genetic inference is that they contain all the information about the history of each region of the genome in a sample of sequences. Therefore, methods that employ ARGs are utilizing the maximum possible information present in the data. The inference of ARGs is made difficult by the large state space of possible graph topologies and the relatively small amount of information mutations provide about the underlying graph. However, recent methods to infer ARGs have made significant progress on this difficult computational task (Rasmussen et al. 2014; Kelleher et al. 2019; Speidel et al. 2019; Mahmoudi et al. 2022; Zhang et al. 2023; Deng et al. 2024). Nevertheless, there will still always be uncertainty as to the underlying ARG, as there will be many ARGs that are compatible with the observed sequence data. For that reason, CLUES also integrates over the uncertainty in the ARG estimation, in addition to integrating over the set of possible allele histories, thus giving two sets of latent variables that are integrated out to obtain the full-likelihood function.

We here present CLUES2, an extension of the CLUES method, which is able to work not only on importance samples of ARGs on modern data but also on ancient genotype samples and ARGs built on ancient data, thus enabling the usage of time series data and linked SNPs to give better estimates of the selection coefficient of an allele. We also develop and test the capability of CLUES2 to estimate different selection coefficients in different time periods. In addition, we make significant improvements to runtime by making several well-justified approximations to the forward and backward algorithms. The computational speedups allow more replicates to be used in the importance sampler, which increases accuracy and also allows CLUES2 to be used in genome-wide scans of selection, where hundreds of thousands or millions of SNPs may need to be analyzed. We perform rigorous testing of both selection coefficient estimation and the appropriateness of the χ^2 test for hypothesis testing of selection. Finally, we improve the interpretability of our reconstruction of historic allele frequencies by integrating over the uncertainty in the estimation of the selection coefficients. We make CLUES2 available on GitHub as a well-documented Python package at https://github.com/avaughn271/CLUES2.

Materials and Methods

CLUES2 Framework

We begin with a short review of the framework of the original CLUES method. CLUES seeks to compute the likelihood of sequence data D given a selection coefficient s. It does this by integrating over the uncertainty in the historic trajectory of the allele under consideration, denoted X, and the uncertainty in the gene tree at the locus of interest, denoted T. Conceptually, CLUES calculates $P(D \mid s)$ as $\int_T \int_X P(D, T, X \mid s)\, dX\, dT$. Concretely, the integration over allele trajectories X is done via the forward algorithm in a hidden Markov model (HMM) framework and the integration over gene trees is done via importance sampling. There are, therefore, two latent variables that are integrated out to generate the estimate of the likelihood function. CLUES2 keeps the same HMM and importance sampling framework to integrate over these two latent variables, although the specifics of the implementation differ. We also extend the capability of CLUES2 to consider a broader set of emissions, allow for more general parameterizations of s, and make substantial improvements to computational complexity.

We devote the rest of this section to describing the HMM of CLUES2. The framework of this model is the same as in CLUES, but the exact expressions for the transition probabilities and the way allele frequencies are discretized differ. We analyze a derived allele that arose in a single copy at an unspecified time in the past. We assume the infinite sites model (Kimura 1969), which implies that there is no back mutation, no recurrent mutation, and that the SNP is biallelic. We wish to find the value of s, the selection coefficient of the derived allele, that maximizes the likelihood of the data D. What exactly D represents depends on the kind of data being analyzed, and we postpone a formal definition of D until the subsequent sections. We compute this likelihood by conditioning on the historic trajectory of the derived allele frequency. In particular, if we let $\mathcal{X}$ be the set of all possible derived allele trajectories, we compute the likelihood as

$$P(D \mid s) = \sum_{X \in \mathcal{X}} P(D, X \mid s) = \sum_{X \in \mathcal{X}} P(D \mid X, s) P(X \mid s)$$
$$= \sum_{X \in \mathcal{X}} P(D \mid X) P(X \mid s)$$

where the last equality follows from the fact that the data is independent of the selection coefficient given the allele trajectory. If we discretize the derived allele frequency into K frequency bins and consider the history of an allele until a maximum time point T (measured in discrete units of 1 generation), computing this expression would naively require summing over K^T many allele trajectories. However, we instead recognize this as a HMM where the derived allele frequency at a given time is the hidden state and the data are the emissions. We can then efficiently compute this sum using the forward algorithm (Baum 1972). This modeling of allele frequencies as hidden states of an HMM has previously been applied by Williamson and Slatkin (1999), Bollback et al. (2008), Steinrücken et al. (2014), Bergman et al. (2018), and Paris et al. (2019). More details of this framework and the underlying population genetic assumptions are given in section "Population genetic models", Supplementary Material online.

The implementation of the forward algorithm is as follows: we compute a forward algorithm matrix F, which has dimension $K \times T$. The forward algorithm evolves backward in time, meaning that the first column of F corresponds to the present, (which is to say 0 generations before the present) and the tth column of F corresponds to $t-1$ generations before the present. Each entry $F_{k,1}$ in the first column is initialized to 0, except that a 1 is placed in the row corresponding to the allele frequency bin closest to the observed modern allele frequency. Assuming the first $t-1$ columns of F have been computed, one timestep of the forward algorithm consists of computing, for each frequency bin k,

$$F_{k,t} = e(t,\, k) \sum_{l=1}^{K} F_{l,t-1} P_{l,k}$$

where $P_{l,k}$ is the probability of transitioning from allele frequency bin l to bin k in one generation and $e(t, k)$ is the probability of observing the data at time t, given that the state is k. To define K frequency bins, we consider the K numbers that are equally spaced between 0 and 1, inclusive, call them $w_1, \ldots, w_K$. We set the allele frequency of bin k, call it x_k, to be the quantile function of a Beta(1/2,1/2) distribution at w_k. This creates a spacing of numbers between 0 and 1, inclusive, that is denser near the boundaries. We model additive selection, meaning the relative fitnesses of ancestral allele homozygotes, heterozygotes, and derived allele homozygotes are 1, $1 + s/2$, and $1 + s$, respectively (we consider extensions of this model in section "Selection with Dominance", Supplementary Material online). We let $\Phi_k(y)$ be the cumulative distribution function (CDF) of a normal distribution with mean $x_k + \frac{s x_k (x_k - 1)}{2 s x_k + 2}$ and variance $\frac{x_k(1-x_k)}{N}$, evaluated at the point y (we omit the dependence of Φ_k on s and N for brevity). Then, for $1 < l < K$, $P_{k,l}$ is calculated as $\Phi_k(\frac{x_l + x_{l+1}}{2}) - \Phi_k(\frac{x_{l-1} + x_l}{2})$, where $\Phi_{s,N}(x)$. $P_{k,1}$ is $\Phi_k(\frac{x_1 + x_2}{2})$ and $P_{k,K} = 1 - \Phi_k(\frac{x_{K-1} + x_K}{2})$ (see section "Allele Frequency Transition Probabilities Under Dominance", Supplementary Material online for further discussion of the derivation of these transition probabilities). We note that the expression for the variance in allele frequency changes implies that near the boundaries of the state space (allele frequency 0 and 1), the changes in allele frequency are expected to be quite small. It is the desire to capture these small fluctuations in allele frequency that motivates our denser spacing of frequency bins near the boundaries. It follows directly from the transition probability definitions that frequencies 0 and 1 are absorbing states. The quantiles of the Beta(1/2,1/2) are chosen because empirically they gives a good approximation, but any spacing scheme would work equally well in the limit as K approaches infinity (provided the distance between adjacent bins approaches 0). Furthermore, it is worth highlighting that throughout this manuscript, we use N to refer to haploid effective population size, denoting the number of sequences present in a population, not the number of diploid individuals in that population. After computing each entry of the matrix F, we can compute the likelihood as $P(D \mid s) = \sum_{k=1}^{K} F_{k,T}$, as we assume uniform exit probabilities from the chain. We then optimize $P(D \mid s)$ with respect to s to achieve our maximum likelihood estimate (MLE)$\hat{s}^{\text{MLE}}$. We use Brent's Method on the interval $[-0.1, 0.1]$. Dividing by $P(D \mid s = 0)$ yields the log-likelihood ratio, which can then be used for hypothesis testing (see the section "Validation of χ^2 Test"). In the following sections, we discuss the types of data we use for the emissions $e(t, k)$.

Ancient Genotypes

One type of emissions considered by CLUES2 are ancient genotypes, which could not be incorporated into the original CLUES method. This is a list of genotypes of ancient individuals along with the times at which these individuals were sampled. If we denote a homozygous derived genotype as DD, a homozygous ancestral genotype as AA and a heterozygous genotype as AD we have

$$e(t,\, k) = \prod_{g \in G_t} \left(x_k^2 \mathbf{1}_{\{g=DD\}} + 2x_k(1-x_k)\mathbf{1}_{\{g=AD\}} + (1-x_k)^2 \mathbf{1}_{\{g=AA\}} \right)$$

where G_t represents the set of all genotypes sampled at time t. This is equivalent to sampling from a binomial distribution with two trials and success probability x_k. Note that it is assumed that each individual is drawn uniformly at random from the population, meaning factors such as relatedness between sampled individuals or a correlation between genotype and likelihood of being sampled could cause bias.

In practice, we allow the usage of genotype likelihoods, so our expression becomes

$$e(t,\, k) = \prod_{g \in G_t} \left(x_k^2 P(R \mid g = DD) + 2x_k(1-x_k) P(R \mid g = AD) + (1-x_k)^2 P(R \mid g = AA) \right)$$

where here R denotes the sequencing read data. We simulate ancient genotype data (as described in section "Ancient Genotypes", Supplementary Material online) and validate

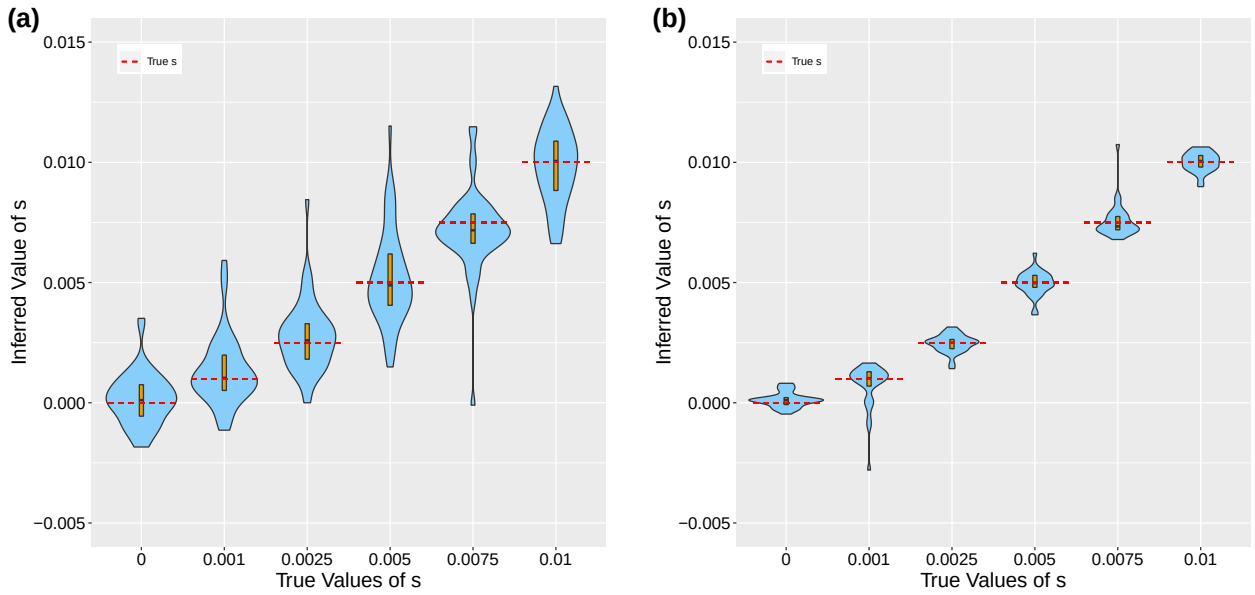

Fig. 1. Violin plots showing the results of running CLUES2 on ancient genotype data. Boxplots are overlayed with the whiskers omitted. True values of s are shown as dashed lines. Thirty replicates were performed for each true value of s. Simulations were run with a) $N = 50{,}000$ and two individuals sampled every generation and with b) $N = 6{,}00{,}000$ and ten individuals sampled every generation.

this approach on the simulated data (see Fig. 1a). If our implementation is correct, there should be exactly two sources of error, which both contribute to variance of our estimator:

1) Fluctuations in the true allele frequency trajectory due to genetic drift, which causes the true allele frequency to be less informative of the value of s.
2) Noisy estimation of the true allele frequency due to finite sampling of ancient genotypes.

Therefore, we ran simulations with very low drift and very dense sampling of ancient genotypes (see Fig. 1b). We observed that the estimated values of $\hat{s}^{\text{MLE}}$ approximately converged to the true values, which validates the correctness of our implementation. We also allow for the usage of ancient haplotype likelihoods, in addition to genotype likelihoods, by substituting the expression $x_k P(R \mid g = D) + (1 - x_k)P(R \mid g = A)$ in the product above.

True Trees

The next class of emissions we consider are coalescence events of a gene tree. While these emissions were also incorporated into the original CLUES method, our exact computation of the coalescent emissions at each timestep, and the handling of the lineage on which the mutation arose, differ. As in the original CLUES paper, we consider a structured coalescent model of the gene tree of a locus under selection following Kaplan et al. (1988) and Braverman et al. (1995). In particular, we assume we know the allele labeling at each leaf node of the tree and that the gene tree satisfies the infinite sites assumption with respect to the allele (we discuss violations to the infinite sites assumption later in the "Inferring Gene Trees on Ancient Human Data" section). We can then obtain a labeling of each branch in the tree as a derived allele branch or an ancestral branch and a labeling of each coalescence node in the tree as a derived node or an ancestral node (see Fig. 2). The one exception to this is the branch on which the mutation must have arisen, which we call the mixed lineage.

Given our labeling, we begin at the present and move back in time, keeping track of the number of derived lineages n_D and ancestral lineages n_A that are present at the given time point. As a bookkeeping measure, n_D also includes the mixed lineage, but we do properly treat this lineage differently, as explained later in this section. At time points younger than the mixed coalescence node, lineages can only coalesce with other lineages of the same allelic class. The instantaneous coalescence rate within the derived class is $X_t N$, and the instantaneous coalescence rate within the ancestral class is $(1 - X_t)N$, where X_t is the frequency of the derived allele at time t. At time points older than the age of the mixed lineage, after which there can be no derived lineages, we only consider coalescence within the ancestral class. A similar structured coalescent approach is used by the simulation softwares discoal (Kern and Schrider 2016) and msprime 1.0 (Baumdicker et al. 2021).

To formalize the emissions in this model, given the model enters time t with n_D derived lineages and d derived coalescences are observed in the interval $(t, t + 1)$ at times t_1 to t_d (where we again emphasize that t increases backwards in time), the derived coalescence emission for frequency bin k is computed as

$$e_D(t, k) = \left(1 - F\left(\frac{\binom{n_D - d}{2}}{x_k N}, t + 1 - t_d\right)\right) \times \prod_{i=1}^{d} f\left(\frac{\binom{n_D - i + 1}{2}}{x_k N}, t_i - t_{i-1}\right)$$

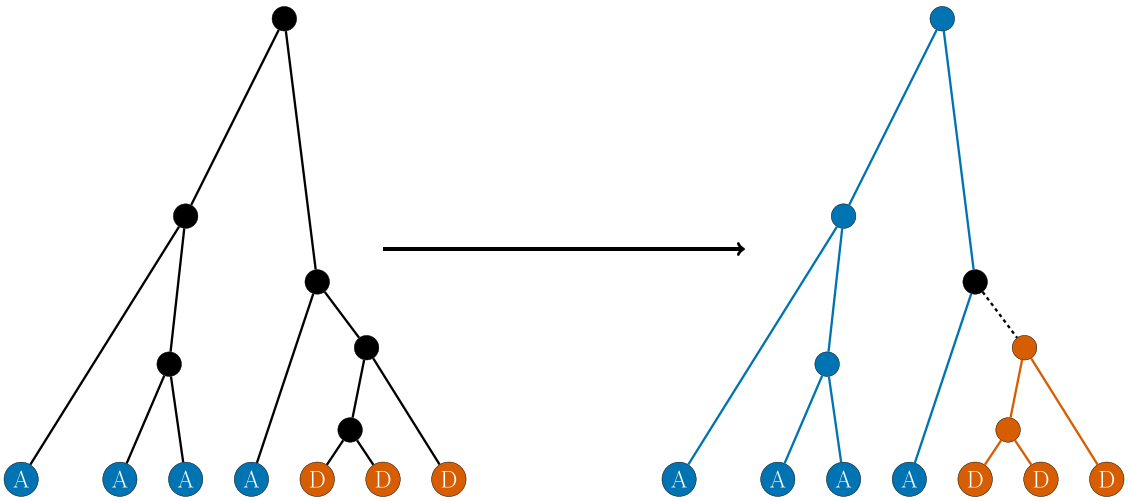

Fig. 2. An outline of the labeling of branches and nodes in a tree as either ancestral (blue) or derived (orange) given a labeling of the leaf nodes. An internal node is a derived coalescence if and only if all its descendant leaves are derived. The parent node of the oldest derived coalescence is the mixed coalescence node (represented in black). All other coalescence events are ancestral coalescences. A branch represents a derived lineage if and only if it has a derived node as an ancestor. The mixed lineage is the immediate parent branch of the oldest derived coalescence (black dashed line). All other branches represent ancestral lineages.

where $f(\lambda, x)$ is the density function of an $Exp(\lambda)$ random variable at the point x, $F(\lambda, x)$ is the corresponding CDF, and $t_0 = t$. This is the probability of observing the given coalescences (possibly none) at the observed times and then not observing another coalescence before the end of the interval $(t, t+1)$. We compute the ancestral coalescence emissions $e_A(t, k)$ in the same way, replacing $x_k N$ with $(1 - x_k)N$ and considering the number of ancestral lineages n_A and the number of ancestral coalescences a in the interval $(t, t+1)$, meaning in total our emission for this timestep is $e(t, k) = e_D(t, k)e_A(t, k)$. Note that we use a continuous-time formulation of the coalescent, as it is only the derived allele frequency that is discretized across timesteps. We also assume that each leaf is drawn uniformly at random and independently from the population.

The exceptions to the above expression are the following:

- If $n_A = 1$ and $n_D = 0$: $e(t, k) = 1$ if $k = 1$ and $e(t, k) = 0$ otherwise.
- If $n_A > 1$ and $n_D = 0$: $e_A(t, k)$ is calculated as usual if $k = 1$ and $e_A(t, k) = 0$ otherwise.
- If $n_A = 1$ and $n_D = 1$: an ancestral emission with $n_A = 2$ is emitted if $k = 1$ (signaling that the derived allele is no longer segregating) and $e(t, k) = 1$ otherwise.
- If $n_A > 1$ and $n_D = 1$: an ancestral emission is emitted as usual if $k \neq 1$ and an ancestral emission with $n_A + 1$ is emitted if $k = 1$.

We simulate true gene trees under selection with msprime (Baumdicker et al. 2021), as described in the section "True Trees", Supplementary Material online, and run CLUES2 on the simulated trees. We find that this framework models the mutation origin and the mixed lineage properly (Fig. 3a). We note that the exact handling of the mixed lineage and the formula for the emissions produced by coalescences differs from that described in the original CLUES method.

There are again two sources of statistical noise in this estimation, which both contribute to the variance

1) Fluctuations in the true allele frequency trajectory due to genetic drift, which causes the true allele frequency to be less informative of the true value of s.
2) Noisy estimation of the true allele history due to the finite number of lineages sampled.

Therefore, we ran simulations with very low drift and a very high number of sampled leaves (Fig. 3b). We observed that the estimated values of $\hat{s}^{\mathrm{MLE}}$ converged to the true values, which indicates the correctness of our implementation and that the model assumptions we make (such as discretizing time and allele frequency) are sufficiently accurate to not cause substantial biases.

Importance Sampling of Trees

Of course, the true topology is never known with 100% certainty for real sequence data. Rather, the observed data consists of a set of SNPs around a locus of interest, meaning that the ARG must be estimated and one must integrate over samples of the tree at the locus of interest in order to obtain an estimate of the likelihood function. In theory, this likelihood could be calculated as

$$\begin{aligned} P(D \mid s) &= \int_G P(D, G \mid s) dG \\ &= \int_G P(D \mid G, s) P(G \mid s) dG = E_{G \mid s}[P(D \mid G)] \\ &\approx \frac{1}{M} \sum_{m=1}^{M} P(D \mid G^m) \end{aligned}$$

where each graph G^m is sampled from the distribution of ARGs given s. However, this would require resampling

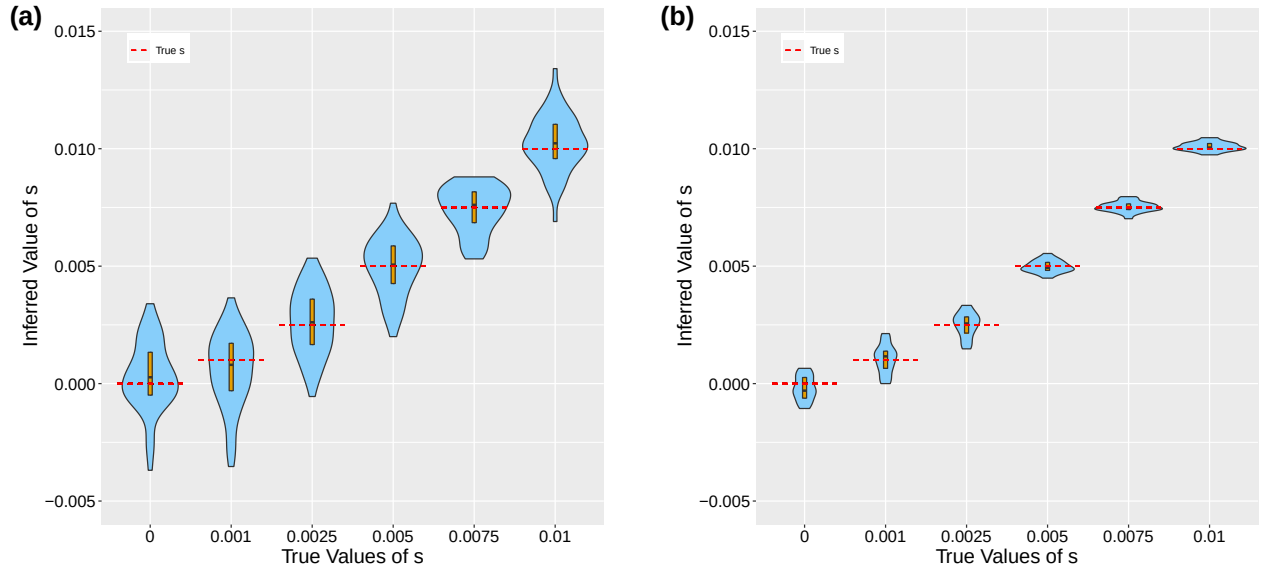

Fig. 3. Violin plots showing the results of running CLUES2 on true trees. Boxplots are overlayed with the whiskers omitted. True values of s are shown as dashed lines. Thirty replicates were performed for each true value of s. Simulations were run with a) $N = 30{,}000$ and 120 sampled leaves and b) $N = 6{,}00{,}000$ and 800 samples leaves. A modern allele frequency of 0.75 was used for each simulation.

graphs every time $P(D \mid s)$ must be evaluated at a new value of s, which is computationally inefficient. To this end, we follow the approach described in the original CLUES method, specifically Equation 18 of Stern et al. (2019), which showed that under the assumption that the ARG G is independent of s given the marginal tree G_k at the SNP of interest, the likelihood ratio $\frac{P(D \mid s)}{P(D \mid s=0)}$ is equal to

$$E_{G \mid D, s=0}\left[\frac{P(G_k \mid s)}{P(G_k \mid s=0)}\right]$$

Therefore, by sampling M graphs $G^1, \ldots, G^M$ from the distribution of ARGs given D and given $s = 0$ and extracting the marginal tree at our SNP of interest from each graph, one can obtain the following Monte Carlo estimator of the likelihood ratio

$$\frac{P(D \mid s)}{P(D \mid s=0)} \approx \frac{1}{M}\sum_{m=1}^{M}\frac{P(G_k^m \mid s)}{P(G_k^m \mid s=0)} \tag{1}$$

which can be computed by sampling many trees and, for each sampled tree, computing the probability of that tree using the expression derived in the previous section. This reweighting of variates sampled from a different distribution is a technique known as *importance sampling* and importantly, does not require resampling graphs every time the likelihood function needs to be evaluated at a new point. We also highlight that the likelihood ratio $P(D \mid s)/P(D \mid s=0)$ is directly proportional to the likelihood $P(D \mid s)$, meaning that the value of s that maximizes the likelihood ratio will be $\hat{s}^{\text{MLE}}$. It is important to note that this approach adds two additional sources of error, in addition to those described previously.

(1) Lack of convergence of the Monte Carlo estimator due to finite M.
(2) Error incurred by improper sampling of ARGs from the specified posterior distribution. This happens when the sampled graphs $G^1, \ldots, G^M$ are not *iid* samples from the distribution $P(G \mid D, s = 0)$. This can happen due to practical considerations, such as poor mixing of the Markov chain Monte Carlo (MCMC) method used to generate these samples, or poor calibration of the ARG-inference methods themselves.

In particular, with regards to the second possible source of error, we highlight the work of Brandt et al. (2022) in showing the degree of miscalibration of different ARG-inference methods and how this depends on parameters such as recombination and mutation rate. Therefore, we develop our own MCMC algorithm for sampling gene trees on sequence data in the absence of recombination (see the section "Importance Sampling", Supplementary Material online for more information). This enables us to validate the correctness of our importance sampling implementation without being affected by possible biases induced by improper sampling of ARG-inference methods. We demonstrate the behavior of our importance sampling estimator through a set of illustrative simulations. We begin by simulating genetic data in msprime on a set of 24 haplotypes for a 1 Mb region with no recombination and with a mutation rate of 3×10^{-6}. We then run our purpose-built MCMC sampler to generate samples of gene trees given the observed genetic data and use these samples as our importance sampling replicates $G_k^1, \ldots, G_k^M$ in Equation (1). We take one sample of a tree ($M = 1$) and use it as input to CLUES2. This corresponds to not using

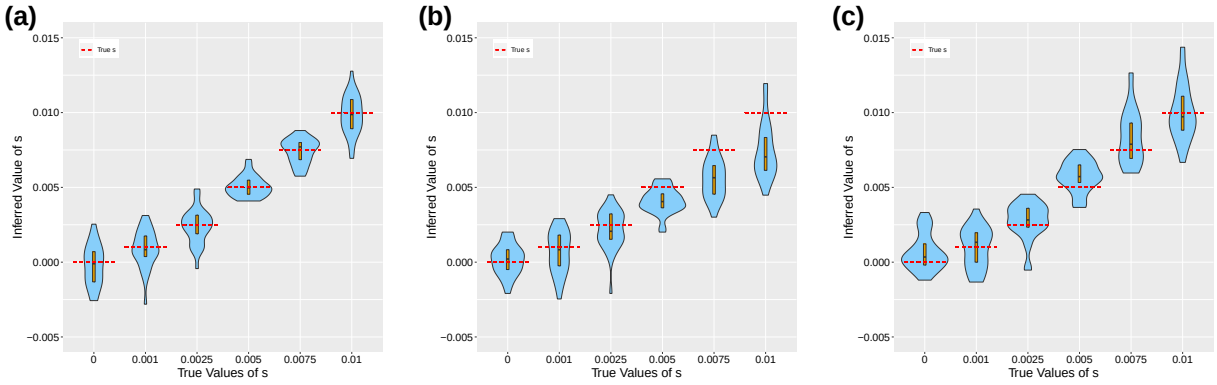

Fig. 4. Violin plots showing the results of running CLUES2 on inferred topologies. Boxplots are overlayed with the whiskers omitted. True values of s are shown as dashed lines. Thirty replicates were performed for each true value of s. Simulations were run with a) $\mu = 3 \times 10^{-6}$ and one sample taken without importance sampling, b) $\mu = 4 \times 10^{-9}$ and one sample taken without importance sampling, and c) $\mu = 4 \times 10^{-9}$ and 600 samples taken and used in the importance sampling framework. A modern allele frequency of 0.75, a population size of $N = 40{,}000$, and 24 leaves were used for each simulation.

the importance sampling framework at all but instead regarding this one sampled tree as the true gene tree at this locus with no uncertainty. We find that the mutation rate is so high that it overwhelms the prior centered on $s = 0$ and this one sample of a tree does indeed behave like the true gene tree for this data in terms of accuracy (Fig. 4a). Then, we run simulations under identical parameter settings (including using $M = 1$ with no importance sampling) except that a mutation rate of $\mu = 4 \times 10^{-9}$ is used. In this case, the estimation of s is biased as there is not enough data to overwhelm the prior centered at $s = 0$. This effect is particularly pronounced for larger true values of s, which are the values that differ the most from the prior (Fig. 4b). This bias can be considered an extreme case of the estimation error that can happen due to small M, and it is important to note that this results in bias, rather than increased variance. We note that this bias towards $s = 0$ for small values of M, in addition to poor calibration of existing ARG-inference methods, could explain previous analyses that showed a bias of CLUES towards $s = 0$ (see Fig. 4 of Hejase et al. 2021, Fig. 2 of Temple et al. 2023). We also note that using a value of K, the number of allele frequency bins, that is too small can result in bias towards $s = 0$.

To show the utility of our importance sampling approach, we run the same simulation as before with $\mu = 4 \times 10^{-9}$ but instead take $M = 600$ samples to be used in CLUES2. We find that by properly reweighting the samples of gene trees through our importance sampling framework, we recover approximately unbiased estimates of the selection coefficients (Fig. 4c), which validates the implementation of our importance sampling approach.

Ancient ARGs

Recently, there has been significant interest in using the increasing number of high-quality ancient genomes to infer historic patterns of selection. However, while the usage of time series data has the potential to greatly improve estimates of selection, most existing methods developed for eukaryotes only analyze historic genotype data and do not incorporate information from linked SNPs (Malaspinas et al. 2012; Le et al. 2022; Mathieson and Terhorst 2022) (though we note that linkage has previously been applied to time series data for viruses and bacteria, such as in Illingworth and Mustonen 2011; Illingworth et al. 2014; Terhorst et al. 2015; Sohail et al. 2020, 2022). However, in order to fully utilize all the information present in available ancient data, it would be desirable to have a method that can incorporate linkage information from SNPs around a locus of interest. To this end, we implement the usage of ARGs on ancient genomes in CLUES2 to incorporate both temporal data and linkage data around the focal site. The general HMM framework remains the same, but the possible emissions of the algorithm change. The input data now consists not only of a list of derived coalescence times and ancestral coalescence times, but also a list of sampling times of derived leaves and a list of sampling times of ancestral leaves. We here derive what the emissions are for this model by considering the joint probability of an observed tree and a set of ancient allele observations given an allele frequency trajectory X. Let τ_D and τ_A be the coalescence times of the derived and ancestral lineages, respectively, and let S_{ancient} be the allelic states of the ancient samples. We assume the sampling times are fixed, known, and are implicitly conditioned upon. We also assume that the current allele frequency, f, is known, i.e. that the sample size of the modern sample is so large that it is not necessary to model uncertainty in f. f is embedded in X. Then

$$\begin{aligned} P(\text{Tree}, S_{\text{ancient}} \mid X) &\propto P(\tau_D, \tau_A, S_{\text{ancient}} \mid X) \\ &= P(\tau_D, \tau_A, \mid X) P(S_{\text{ancient}} \mid X) \end{aligned}$$

because the tree is fully determined by τ_D and τ_A up to a combinatorial term that describes the possible topologies compatible with τ_D, τ_A and which, importantly, does not depend on X.

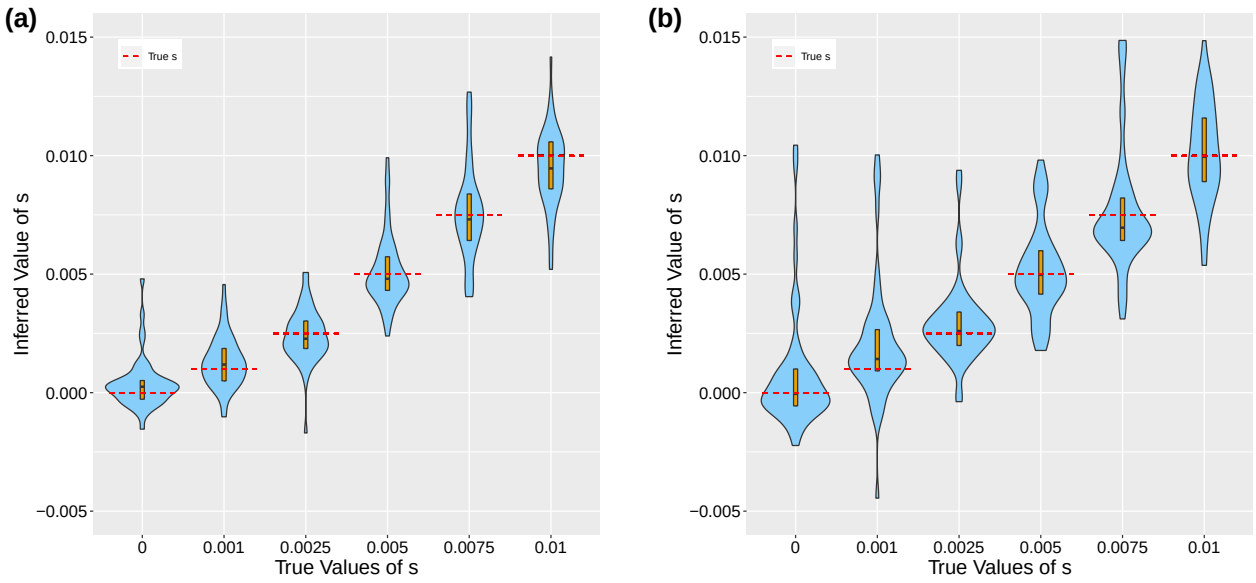

Fig. 5. Violin plots showing the results of running CLUES2 on a combination of modern and ancient data. Boxplots are overlayed with the whiskers omitted. True values of s are shown as dashed lines. Fifty replicates were performed, with each replicate generating an estimate for each of the three selection coefficients. We run simulations a) where the ancient data is incorporated into the tree and b) where the ancient data is treated only as genotype emissions. $N = 40{,}000$, 20 modern leaves are used, and 80 ancient leaves or 40 ancient genotypes are sampled at each of the times 50 and 100 generations before the present.

This derivation implies the following set of emissions. If, for a given time t, there are no leaves sampled in the interval $(t, t+1)$, then the emissions are the same as for the case of modern ARGs, $e(t, k) = e_D(t, k)e_A(t, k)$. However, if we sample m_D derived leaves in this interval and m_A ancestral leaves in this interval, then our emission becomes $e(t, k) = e_D(t, k)e_A(t, k)x_k^{m_D}(1-x_k)^{m_A}$ to account for the genotype observations of the leaves. Note that we assume that genotypes are hard-called and do not allow the usage of genotype likelihoods for the leaves of the ARG. After each timestep, we also increase the current number of derived lineages and ancestral lineages as necessary, corresponding to any sampled leaves. To illustrate the correctness of this approach, we simulate ARGs on ancient data under different selection coefficients (see the section "Gene Trees on Ancient Data", Supplementary Material online for our simulation methodology) and find that we obtain approximately unbiased estimates of the selection coefficient (Fig. 5a). Note that while these specific simulations use the true gene trees, importance sampling on ancient ARGs is also possible. To show the improvements in accuracy obtained by utilizing this approach, we also ran CLUES2 using only the modern ARGs obtained from these data and utilizing the ancient data only as genotype emissions, rather than incorporating them into the ARG (Fig. 5b). We see that while the estimates of selection are still approximately unbiased, the variance in the estimation is larger due to the smaller amount of linkage information being used in the analysis. Therefore, we highlight the ability of this approach to utilize both time series and linkage information to generate the most accurate possible estimates of selection coefficients.

Selection in Multiple Epochs

We also develop new functionality in CLUES2 for the joint inference of selection coefficients that differ between time periods. This is done by changing the mean of the normal distribution for the transition probabilities from $x_k + \frac{sx_k(x_k-1)}{2sx_k+2}$ to $x_k + \frac{s_t x_k(x_k-1)}{2s_t x_k+2}$ where s_t depends on the timestep t. In particular, we allow time breakpoints $\tau_1, \ldots, \tau_n$ to be specified, which results in $n+1$ different selection coefficients being fit, one for each of the epochs $[0, \tau_1)$, $[\tau_1, \tau_2)$,..., $[\tau_n, T)$. We now use Nelder–Mead for our optimization method. If we estimate selection in M epochs, our initial simplex is defined by the set of M points for which $s = 0.01$ in one epoch and $s = 0$ in all other epochs in addition to the point defined by $s = 0$ in all epochs, giving $M+1$ points total. To validate our approach, we simulated ancient genotype data from a model where the selection coefficient in the epoch [0, 200) was 0.01, the selection coefficient in the epoch [200, 600) was 0 and the selection coefficient in the epoch [600, inf) was −0.005. We find that we are accurately able to jointly infer these three different selection coefficients (Fig. 6).

Reconstructing Historic Allele Frequencies

In addition to the inference of selection coefficients and log-likelihood ratios of selection, CLUES2 also has the capability to reconstruct historic allele frequency trajectories. Specifically, we calculate, for each time t, the posterior distribution over allele frequencies given the estimate of s and the data D. Note that this is the marginal posterior distribution of allele frequencies at a particular time, which is different from considering the joint distribution of allele

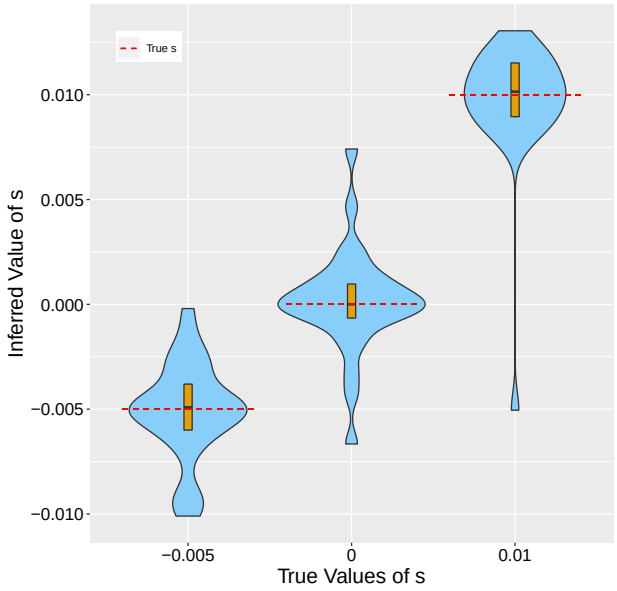

Fig. 6. Violin plots showing the results of running CLUES2 on ancient genotype data simulated with differing selection coefficients through time. Boxplots are overlayed with the whiskers omitted. True values of s are shown as dashed lines. Thirty replicates were performed, with each replicate generating an estimate for each of the three selection coefficients. A population size of $N = 70{,}000$ was used and eight diploid individuals were sampled in each generation.

frequencies across all times. This is accomplished by running the backward algorithm (Baum 1972), the complement to the forward algorithm, which consists of calculating a matrix B following the recursion

$$B_{t,k} = \sum_{l=1}^{K} e(t+1, l) P_{k,l} B_{t+1,l}$$

where the last column is initialized to all ones. Then, the desired posterior on allele frequency for a given time point t, call it π_t, has the form $\pi_t(k) \propto F_{t,k} B_{t,k}$.

In the original CLUES algorithm, the full forward–backward algorithm is run with the only value of s considered being the estimate of $\hat{s}^{\text{MLE}}$. However, we find that this does not incorporate the uncertainty inherent to the estimation of s and, therefore, results in posterior distributions that are underdispersed. We instead wish to compute the posterior distribution on allele frequencies $P(X_t = k \mid D)$ assuming a uniform prior on s. The choice of a uniform prior is motivated by our maximum likelihood estimation framework for computing $\hat{s}^{\text{MLE}}$, which could be rephrased as posterior inference with a uniform prior on s. With a uniform prior on s, this distribution is

$$\begin{aligned} P(X_t = k \mid D) &= \int_{-\infty}^{\infty} P(X_t = k, s \mid D) ds \\ &= \int_{-\infty}^{\infty} P(X_t = k \mid s, D) P(s \mid D) ds \\ &\propto \int_{-\infty}^{\infty} P(X_t = k \mid s, D) P(D \mid s) ds \\ &= \int_{-\infty}^{\infty} \pi_t^s(k) P(D \mid s) ds \end{aligned}$$

In theory, one could choose a set of discrete values for s, call them $s_1, \ldots, s_N$, and approximate this integral using Simpson's rule, numerical quadrature, or any other way for deterministically approximating the area under the curve of interest. However, these methods scale poorly to higher-dimensional spaces, which will be the case when we are fitting multiple selection coefficients. For this reason, we choose to use Monte Carlo integration, as the variance in the estimate of the integral does not depend on the dimension of the state space but is instead always proportional to $1/M$ where M denotes the number of points used in the estimation, (see e.g. Jarosz 2008). Our Monte Carlo estimator is

$$\int_{-\infty}^{\infty} \pi_t^s(k) P(D \mid s) ds \approx \frac{1}{M} \sum_{i=1}^{M} \pi_t^{s_i}(k)$$

where the s_i are sampled from $P(D \mid s)$. To do this, we assume that $P(D \mid s)$ is approximately normally distributed with mean equal to $\hat{s}^{\text{MLE}}$ as computed by our optimization algorithm. This approximation is based on the Bernstein–von Mises theorem, which states that asymptotically, the posterior is normally distributed with a variance given by the Fisher information matrix. Therefore, in the limit of large data and many Monte Carlo samples, this approximation becomes exact.

Using the pairs of selection coefficients and values of the likelihood function $P(D \mid s)$ computed during our optimization routine, call them $(s_j, L(s_j))$, we then fit the variance (or covariance matrix) to our data points using a least-squares regression model. Concretely, we find the value of Σ that minimizes

$$\sum_j \left(L(s_j) - \phi(s_j; \hat{s}^{\text{MLE}}, \Sigma) \right)^2$$

where $\phi(x; \mu, \Sigma)$ is the density of a normal distribution with mean μ and covariance matrix Σ evaluated at the point x. We then sample a number of points M from this distribution ($10d$ by default, where d is the dimension of s) and compute $\frac{1}{M} \sum_{i=1}^{M} \pi_t^{s_i}(k)$ as our estimate of the posterior distribution over allele frequencies at time t. In this way, we take uncertainty in our estimation of s into account for our estimate of the posterior distribution of allele frequencies. We validate this approach by comparing it to an exact rejection sampling approach for reconstructing historic allele frequencies. Specifically, we fix a set of genotype observations and run CLUES2 with our Monte Carlo integration approach. Then, we repeatedly sample selection coefficients uniformly from $[-0.1, 0.1]$ and simulate allele histories conditional on our sampled coefficient. We then sample genotype observations conditional on these trajectories and retain only those trajectories associated with datasets that match our fixed set of genotype observations. This is done until we have retained 1,000 trajectories, from which we can reconstruct empirical posterior frequency intervals. We compare these two approaches

in Fig. S5, Supplementary Material online. We also perform a comparison where two selection coefficients are independently inferred, where the coefficient in each interval is sampled uniformly from [−0.1, 0.1] (Fig. S6, Supplementary Material online). For both simulations, we find strong concordance between our Monte Carlo integration framework and the exact rejection sampling approach.

Validation of χ^2 Test

Along with reporting estimates of selection coefficients, we report the log-likelihood ratio $\ln \frac{P(D \mid \hat{s}=s^{\mathrm{MLE}})}{P(D \mid s=0)}$ and aim to compute the associated *P*-value for selection. By Wilks' Theorem, if data are simulated under the null hypothesis of $s = 0$, $2 \ln \frac{P(D \mid \hat{s}=s^{\mathrm{MLE}})}{P(D \mid s=0)}$ should asymptotically follow a χ^2 distribution with degrees of freedom equal to the difference in the dimensions of the parameter spaces of $\hat{s}^{\mathrm{MLE}}$ and $s = 0$ (Wilks 1938). As $s = 0$ has dimension 0, the degrees of freedom is always equal to the number of different selection coefficients that are estimated in different epochs (we set the default value to 1). Furthermore, the *P*-value obtained from this χ^2 distribution should be uniformly distributed on [0, 1] under the null hypothesis.

We, therefore, perform validations on each of our above simulations by examining the distribution of $2 \log(LR)$ when data are simulated under $s = 0$ to the expected χ^2 distribution with k degrees of freedom where k is the number of independent selection coefficients that are estimated (3 in the section "Selection in multiple epochs", 1 in all other simulations). We compare the quantiles of the empirical distribution of $2 \log(LR)$ to the theoretical χ^2 distribution using a P–P plot. We also plot a histogram with five bins of the *P*-values obtained from these log-likelihood ratios by using the upper tail of the appropriate χ^2 distribution and comparing it to the uniform distribution. The results are shown in Figs. S7 to S12, Supplementary Material online. We find that for each simulation we consider, the empirical distribution of $2 \log(LR)$ follows the expected χ^2 distribution and that the corresponding *P*-values are distributed uniformly between 0 and 1, thus indicating that CLUES2 is properly calibrated and can be used in rigorous hypothesis-testing frameworks.

Computational Improvements

We make modifications to the naive forward–backward algorithm that significantly improve the computational runtime of CLUES2. In practice, there are two steps of the basic forward algorithm implementation of CLUES2 that require significant computational runtime. The first is the computation of the $K \times K$ per-generation allele frequency transition matrix P. Naively, this requires the computation of K^2 entries. However, many of these probabilities will be close to 0 as the per-generation probability of transitioning from a high allele frequency to a low allele frequency in one generation and vice-versa is very small. More formally, we observe that the nature of the Gaussian transition probabilities ensures that the transition matrix is a sparse banded matrix where appreciable probability only falls near the main diagonal. Therefore, we instead, for each allele frequency bin k, only compute entries of row k corresponding to frequencies falling within 3.3 SDs of the mean of the Gaussian distribution represented by this row. As 99.9% of the density of a Gaussian density function lies within 3.3 SDs of its mean, we find this captures $\sim 99.9\%$ of the transition probability at each step. We then rescale each row of the matrix to sum to 1. All other entries of the matrix are left at 0. We note that this idea of approximating the full per-generation transition matrix by a sparse banded matrix has been previously applied to the full Wright–Fisher model (Spence et al. 2023). We call this Approximation A1.

The other section that requires significant time is the execution of the forward algorithm, which must be run every time the likelihood function is called during the optimization procedure (or M times for each call to the likelihood function if importance sampling is used with M samples). When looking at the forward algorithm equation:

$$F_{t,k} = e(t, k) \sum_{l=1}^{K} F_{t-1,l} P_{l,k}$$

We notice that most of the terms in the sum will be 0 as we have set most of the entries of the matrix P to 0. This means that these values of l can be left out of the summation without affecting the sum, implying that we instead can compute:

$$F_{t,k} = e(t, k) \sum_{l=a_k}^{b_k} F_{t-1,l} P_{l,k}$$

where a_k and b_k are the lowest and highest indices of the column $P_{\cdot,k}$ that are nonzero. These lower and upper bounds can be computed at the same time as the transition matrix P is computed and do not need to be recomputed at each timestep of the forward algorithm. We call this Approximation A2 to emphasize its dependence on Approximation A1 (although given Approximation A1, this change incurs no additional approximation error).

We also make another computational improvement based on the observation that when computing

$$F_{t,k} = e(t, k) \sum_{l=a_k}^{b_k} F_{t-1,l} P_{l,k}$$

even for indices l with fairly large values of $P_{l,k}$, the computed sum might still be near 0 if the values of $F_{t-1,l}$ are quite small. Therefore, we choose to simply not compute values of $F_{t,k}$ if we can reasonably conclude that they will be near 0. We do this by calculating, at each timestep t, a lower bound α_t and an upper bound β_t between which 99.9% of the probability

of column $F_{t,\cdot}$ lies (i.e. bounds for which $\frac{\sum_{k=\alpha_t}^{\beta_t} F_{t,k}}{\sum_{k=1}^{K} F_{t,k}} > 0.999$). We let x_{α_t} and x_{β_t} be the allele frequencies corresponding to the frequency bins α_t and β_t. We then compute $F_{t+1,k}$ only for a bin k with corresponding frequency x_k satisfying $x_{\alpha_t} - 2(x_{\beta_t} - x_{\alpha_t}) - 0.1 < x_k < x_{\beta_t} + 2(x_{\beta_t} - x_{\alpha_t}) + 0.1$. The other entries in that column of F are left as 0. Informally, we are assuming that each column of F is relatively close in distribution to the previous column of F. This is a reasonable assumption if the emissions at a given timestep do not significantly change the probability distribution across the hidden states. If, for example, the probability distribution across states is concentrated on high derived allele frequency states and at the next timestep 1,000 ancient genotype emissions are observed that are all homozygous for the ancestral allele, then this could result in substantial error in computing F. However, this would imply a huge, one-timestep change in allele frequency, which is incredibly unlikely. As we enforce the fact that the derived allele frequency must be 0 when the mixed lineage coalesces (as described in section "True Trees"), we increase the robustness of this assumption in the following way. If we observe only one derived lineage left at time t, we set α_t to 1, so we always compute $F_{t,1}$ and all values $F_{t,k}$ for $k < \beta_k$ and are, therefore, never "surprised" by observing this coalescence. We find that this approximation saves significant computational time and has a negligible effect on the estimated value of the selection coefficient. This practice of running the forward algorithm while only keeping track of a smaller number of "best" states is inspired by the beam search approaches used to approximate the Viterbi decoding of an HMM (Deshmukh et al. 1999), although the exact details differ. We call this Approximation B. The concepts underlying Approximations A1, A2, and B are shown in Fig. 7. The probability thresholds of 99.9% and the exact frequency bounds of Approximation B are arbitrary and chosen because they were found to significantly speed up the forward algorithm without adversely affecting accuracy.

For the backward algorithm, we also perform Approximation A1, and make an analogous approximation to Approximation A2, where now a_k and b_k are the lowest and highest indices of each row of P (instead of column of P) that are nonzero. We do not make an analogous approximation for Approximation B. In practice, this set of computational improvements greatly improves the runtime of CLUES2. We show this by measuring the runtime of CLUES2 both with and without these approximations as a function of the number of importance samples used for a sample dataset. We plot the results in Fig. 8.

We see that for this dataset, using these approximations results in a speedup of between 1 and 2 orders of magnitude. These computational speedups are important for several reasons. Firstly, they allow more importance samples to be used for the same amount of computational resources, which contributes to better convergence of the Monte Carlo estimator of the log-likelihood ratio (Equation (1)) and thus improved accuracy. Secondly, it expands the applicability of CLUES2 to large numbers of SNPs, for example in genome-wide scans of selection. We show that these approximations have a negligible impact on our estimate of the log-likelihood function by comparing results both with and without these approximations (see Fig. 9).

Inferring Gene Trees on Ancient Human Data

We applied CLUES2 to a recently published collection of imputed and phased ancient genomes (Allentoft et al. 2024). This dataset consists of 1,664 diploid individuals sampled from around the world and has both newly sequenced genomes as well as previously published genomes. The highest concentration of these genomes is located in Western Eurasia (see Fig. 1 of Irving-Pease et al. 2024 for a detailed map of geographic locations and sampling times of these genomes), and for that reason, we restricted our analysis to West Eurasian genomes. In addition, it has been shown that inferring ARGs on very low coverage data can cause biases in the estimation of pairwise coalescent times (see section "Differing Population Sizes", Supplementary Material online of Allentoft et al. 2024), so we chose to limit our analysis to a subset of 187 sampled West Eurasian genomes with a coverage of at least 2×. We then merged this ancient data with a dataset of 100 randomly chosen individuals from the 1,000 Genomes Phase 3 EUR Superpopulation, resulting in 287 total diploid individuals (The 1000 Genomes Project Consortium 2015).

To analyze a chosen set of SNPs, we ran Relate (Speidel et al. 2019) on the corresponding chromosomes of our merged dataset, using recombination maps and genomic masks from the 1,000 Genomes Project Phase 3 in addition to the GRCh37 human ancestral sequence (The 1000 Genomes Project Consortium 2015) to polarize the alleles. The `EstimatePopulationSize` capability was used in order to generate a .coal file representing the estimated historic population sizes, which was then used as input to the main Relate algorithm. The `SampleBranchLengths` function was used to generate 2,000 samples of the marginal tree at each focal SNP, and we converted this output to the CLUES2 input using a custom Python script `RelateToCLUES.py`. For certain SNPs, the sampled topologies did not satisfy the infinite sites assumption with respect to the focal SNP. When this occurs, we performed the minimum number of "leaf flips" such that the infinite sites assumption is satisfied. A "leaf flip" consists of switching the allele of a leaf node from derived to ancestral or vice versa. Note that while the minimum number of flips is always well-defined, the exact leaves to be flipped is not necessarily unique. If this is the case, one set of leaves to flip is chosen deterministically based on the ordering of leaf nodes in the Newick representation generated by Relate via the `SampleBranchLengths` function. See the section "Minimum Leaf Flipping Algorithm", Supplementary Material online for more information on our leaf flipping algorithm. We list the number of leaf flips performed for each SNP in Table 2. We than ran CLUES2 on each of these input files with the arguments `--df 600` (which sets 600

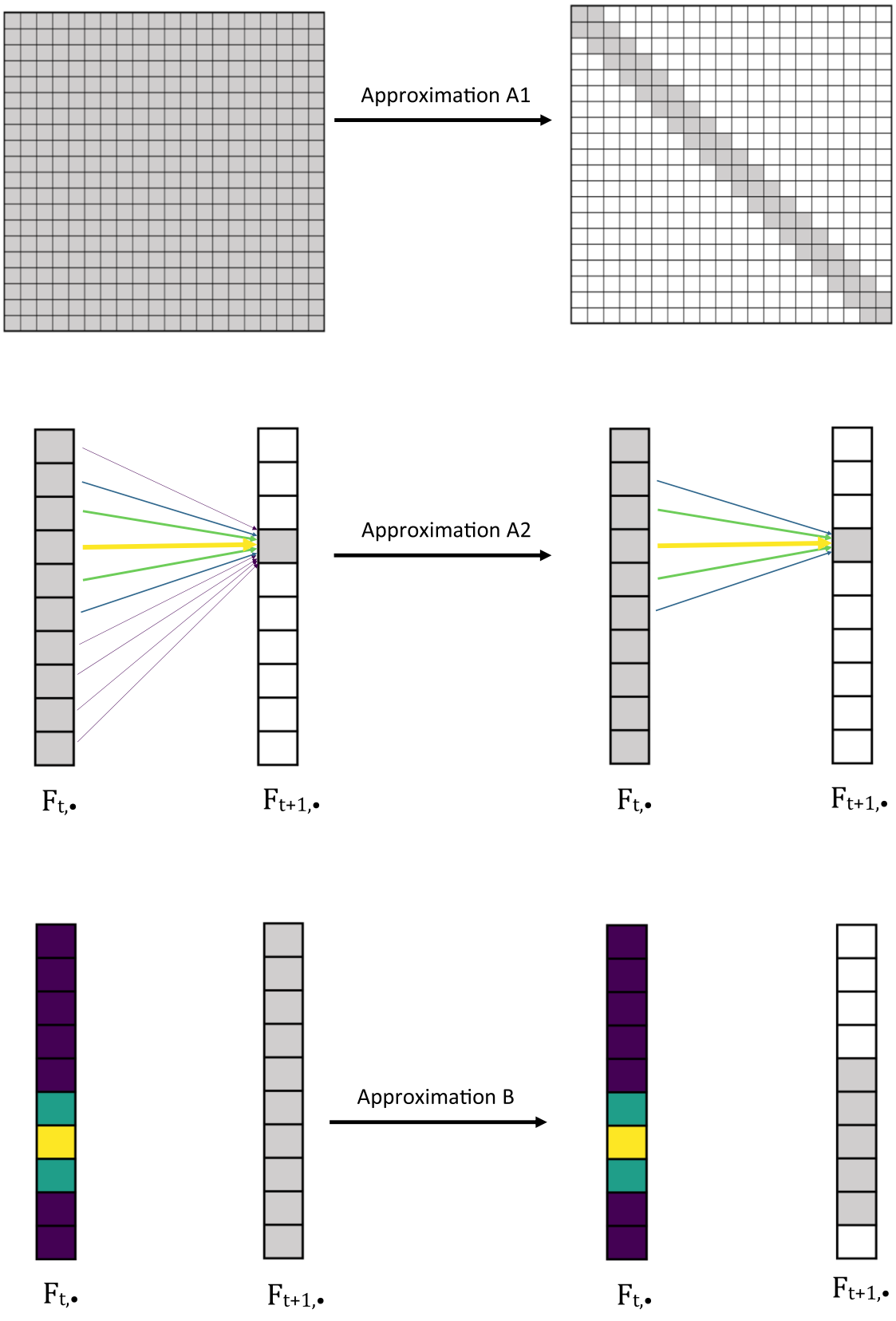

Fig. 7. Illustration of Approximations A1, A2, and B. Approximation A1 approximates the transition matrix by a sparse banded matrix. Approximation A2 reduces the number of states in the previous column of *F* that are summed over to compute each entry of the forward matrix *F*. Approximation B reduces the number of entries that are computed in a column of the forward matrix *F* based on the probability density of the previous column of *F*. Here, the gray entries represent values that are computed, while the colored entries and arrows represent transition or forward probabilities. Lighter colors denote higher probabilities, and darker colors denote smaller probabilities.

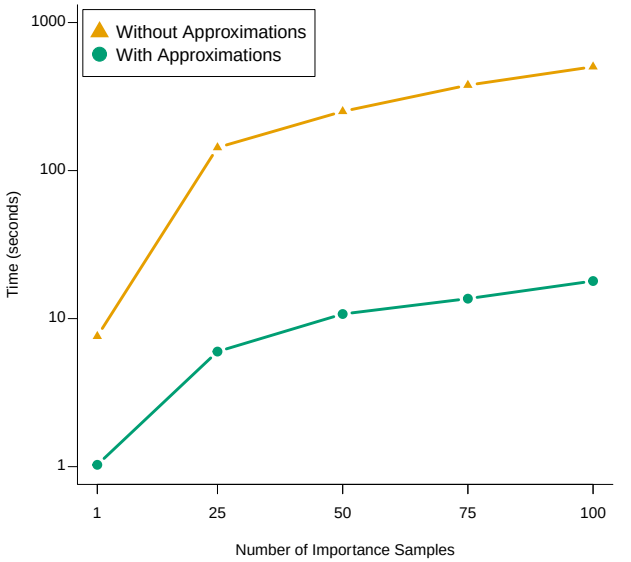

Fig. 8. Comparative runtime of CLUES2 on different numbers of importance samples both with and without the stated approximations.

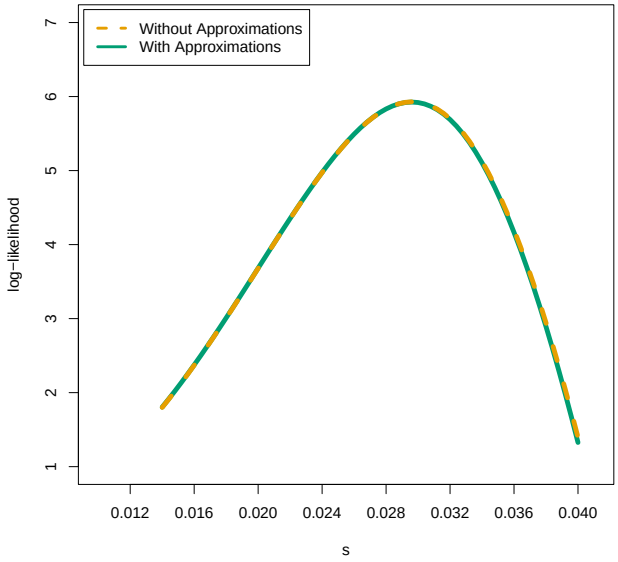

Fig. 9. Comparison of the estimates of the log-likelihood function of our dataset of 100 importance samples both with and without our HMM approximations. The log-likelihood was evaluated at 60 values of s spaced equally between 0.014 and 0.04 for each case, and the plots of the functions were generated via linear interpolation.

discretized allele frequency bins), `--tCutoff 536` (this sets the maximum time of the HMM at 536 generations, which, assuming a generation time of 28 years, (Moorjani et al. 2016) corresponds to 15,008 years), and a value of `--popFreq` (the modern derived allele frequency) that was estimated from our 100 modern individuals.

Results

Comparison with Existing Methods

A variety of existing methods have been used to identify selection coefficients. It is not directly possible to test all features of CLUES2 against these methods as to our knowledge CLUES2 is the first method that can model changing selection coefficients through time in a hypothesis-testing framework and that can do this on samples of ARGs on ancient data. However, we do compare the results of CLUES2 with other methods on our low-mutation simulated genotype data described in the section "Importance Sampling of Trees" and plotted in Fig. 4c. As this data consists of all modern samples and we are only estimating a constant selection coefficient through time, many existing methods are applicable to this data. We consider three other methods, all based on summary statistics of the data. The first is Tajima's D, a site-frequency-spectrum-based statistic which is based on the relative difference between the number of segregating sites and the nucleotide diversity from that expected under neutrality (Tajima 1989). The second is H12, a haplotype-based statistic that calculates the imbalance in the relative frequencies of the different haplotypes observed in the data (Garud et al. 2015). The third is nSL, which is based on calculating the average number of segregating sites around a focal SNP that are identical by state among haplotypes with the derived allele, calculating the corresponding average among haplotypes with the ancestral allele, and taking the log-ratio of the two quantities (Ferrer-Admetlla et al. 2014). We measure the ability of these three methods to infer selection coefficients through an ABC approach, detailed in the following pseudocode.

Algorithm 1 Approximate Bayesian Computation Algorithm

1. **Given dataset D and statistic θ, compute $\theta(D)$**
2. **Initialize m=0 and $S = \{\}$**
3. **While $m < M$:**
   - (a) **Sample $\tilde{s}$ uniformly from the interval (0, 0.025)**
   - (b) **Simulate dataset $\tilde{D}$ with selection coefficient $\tilde{s}$**
   - (c) **Compute $\theta(\tilde{D})$**
   - (d) **If $|\theta(\tilde{D}) - \theta(D)| < \epsilon$:**
     - i. $m \leftarrow m + 1$
     - ii. $S \leftarrow S \cup \{\tilde{s}\}$
4. **Compute empirical density f of samples in S**
5. **Output empirical mode of f**

Notice that we use a uniform (0, 0.025) prior, which allows the empirical mode to be interpreted as an approximation to the maximum likelihood estimator. The datasets we analyze are the same datasets used in the low-mutation simulation described in the section "Importance Sampling", Supplementary Material online, meaning that they consist of 24 haplotypes of a 1 Mb region with a mutation rate of 7×10^{-7}. We run this algorithm for each dataset using $M = 2{,}000$ iterations, and ϵ values of 0.015, 0.001, and 0.05 for Tajima's D, H12, and nSL, respectively. Each summary statistic is computed for the whole region. The posterior is estimated, for each set of resulting ABC sampled values of s (S), using kernel density estimation with the Epanechnikov (parabolic) kernel and a bandwidth specified by Silverman's "rule of thumb" (Epanechnikov 1969; Silverman 1986). We found that increasing M and

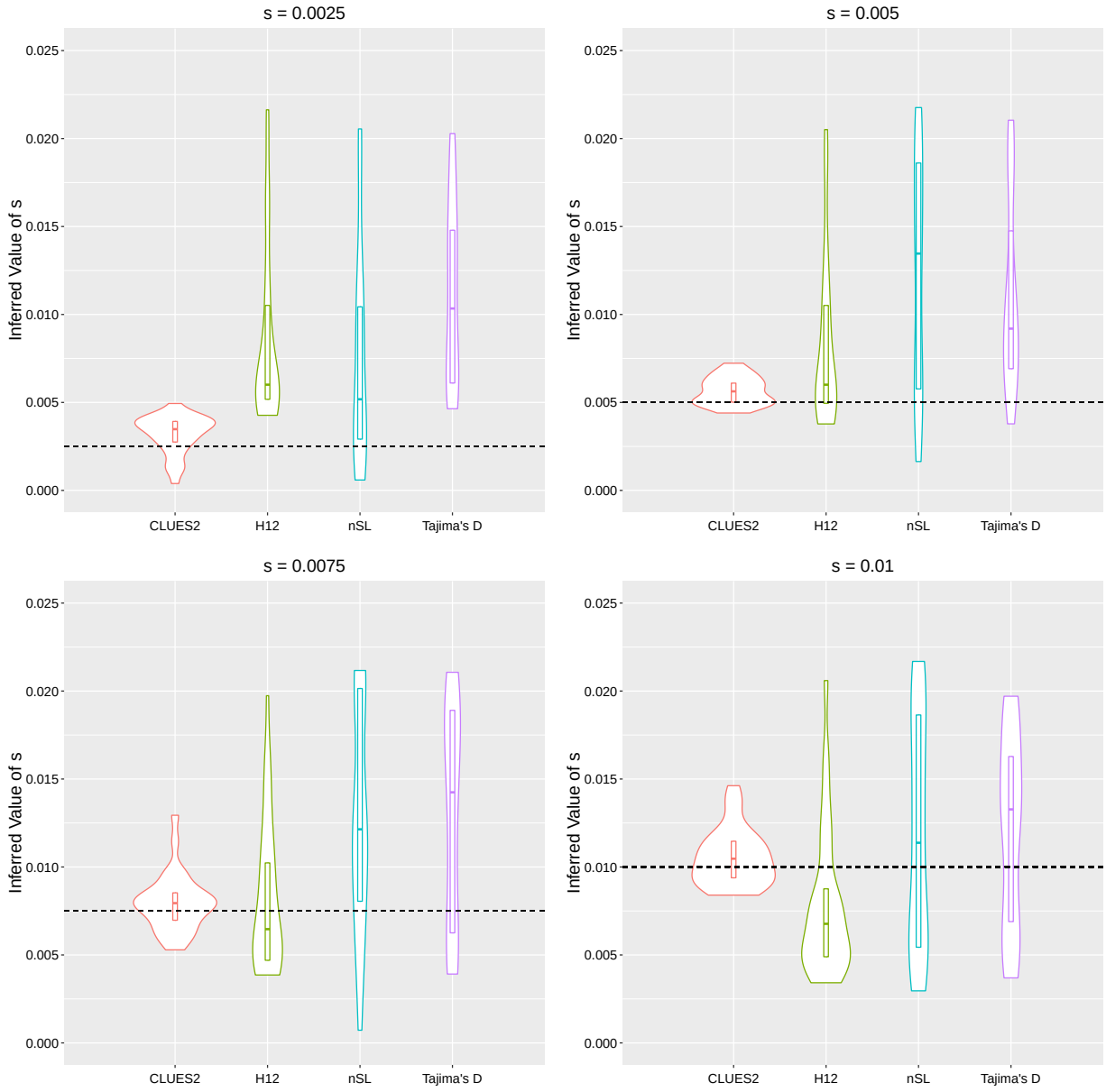

Fig. 10. Comparison of the ability of CLUES2 and several summary statistic-based methods to infer selection coefficients. We infer selection coefficients from the summary statistics using the ABC algorithm described above. The dataset is identical to that analyzed in Fig. 4c. The true simulated values of the selection coefficients are shown as horizontal dashed lines.

decreasing ϵ past these thresholds did not significantly change the estimates. It is worth noting that because of the relatively small set of values H12 can take for 24 haplotypes (it is bounded above by the partition number $p(24) = 1{,}575$), we can explicitly calculate that the value of $\epsilon = 0.001$ reduces this ABC approach to exact rejection sampling. This is to say that two datasets of 24 haplotypes have H12 values that differ by less than 0.001 if and only if their H12 values are exactly the same. We plot the results of this analysis compared with the CLUES2 results in Fig. 10. Note that we only analyze the four largest selection coefficients as `msprime` cannot simulate sweeps with negative selection coefficients, meaning that edge effects in the density estimation can happen for very small selection coefficients.

We see that all the summary statistic-based methods show much greater variance than CLUES2, which is to be expected given the comparative lack of information present in summary statistics when compared with the full dataset.

Analysis of Ancient Human Data

We chose four SNPs to analyze in the ancient human dataset that have previously been identified as candidates for selection in Eurasians. The first is rs4988235, located in the MCM6 gene, where the derived allele is associated with lactase persistence (Enattah et al. 2002; Bersaglieri et al. 2004; Chin et al. 2019). The second is rs35395 in the SLC45A2 gene, where the derived allele is associated with lighter

Table 1 General information about the four SNPs we considered in our ancient human analysis

Gene	SNP rsID	GRCh37 position	Anc/Der allele	Derived allele freq
MCM6	rs4988235	chr2:136608646	G/A	0.535
SLC45A2	rs35395	chr5:33948589	T/C	0.930
TNXB	rs12153855	chr6:32074804	T/C	0.125
ACP2	rs75393320	chr11:47266471	G/C	0.150

Table 2 Results of the one-epoch model. One selection coefficient is estimated that is constant through time

Gene	$s^{MLE}_{0-15,000}$	$-\log_{10}(p)$	AIC	Leaf flips
MCM6	0.01894	59.90	−267.78	2
SLC45A2	0.02650	26.00	−112.51	3
TNXB	0.02062	10.85	−43.66	8
ACP2	0.00869	5.47	−19.56	0

This is the best model for the ACP2 gene, as determined by having the lowest AIC. We report the number of leaf flips necessary for each SNP. We do not repeat this information in the following tables as it depends only on the inferred tree structure and thus remains consistent across models.

Table 3 Results of the two-epoch model

Gene	$s^{MLE}_{0-5,000}$	$s^{MLE}_{5,000-15,000}$	$-\log_{10}(p)$	AIC
MCM6	0.02763	0.00534	64.19	−291.63
SLC45A2	0.02978	0.00260	42.02	−189.49
TNXB	0.03599	−0.01248	13.81	−59.58
ACP2	0.01029	0.00667	4.74	−17.81

There is one selection coefficient breakpoint at 5,000 years before the present. This is the best model for the TNXB gene, as determined by having the lowest AIC.

Table 4 Results of the three-epoch model

Gene	$s^{MLE}_{0-2,500}$	$s^{MLE}_{2,500-5,000}$	$s^{MLE}_{5,000-15,000}$	$-\log_{10}(p)$	AIC
MCM6	0.01570	0.04454	0.00326	64.71	−297.29
SLC45A2	−0.00508	0.07051	0.00066	44.15	−202.22
TNXB	0.04083	0.02690	−0.01086	13.09	−58.00
ACP2	0.00961	0.01134	0.00645	4.15	−15.82

There are two selection coefficient breakpoints at 2,500 and 5,000 years before the present. This is the best model for the MCM6 and SLC45A2 genes, as determined by having the lowest AIC.

skin pigmentation (Marcheco-Teruel et al. 2014; Tiosano et al. 2016; Lloyd-Jones et al. 2017; Lona-Durazo et al. 2019). The third is rs12153855 in the TNXB gene of the human leukocyte antigen (HLA) region, where the derived allele is associated with age-related macular degeneration and atopic dermatitis (Cipriani et al. 2012; Weidinger et al. 2013; Grange et al. 2015; Ye et al. 2016). The fourth is rs75393320 in the ACP2 gene, where the derived allele is associated with increased HDL cholesterol (the so-called "good cholesterol") (Klarin et al. 2018; Ke et al. 2021). General information about each SNP is described in Table 1, where the derived allele frequency is calculated from our 100 sampled EUR individuals from the 1,000 Genomes Project.

For each SNP, we calculated the MLE of the selection coefficient, the *P*-value as computed from the log-likelihood ratio, and the resulting Akaike information criterion (AIC) (Akaike 1974). We ran CLUES2 on each of these SNPs assuming a constant value of the selection coefficients through time, with a selection coefficient breakpoint at 5,000 years ago, and with two breakpoints at 2,500 years ago and 5,000 years ago. We show the results in Tables 2 to 4, where we group the results by the number of selection breakpoints. We also plot our estimates of the derived allele frequency trajectories for each SNP under the model with the lowest AIC in Figs. S1 to S4, Supplementary Material online.

There are several important notes about this analysis. Firstly, we set the selection breakpoints explicitly for each model, rather than estimating each of them. This is done because allowing selection breakpoints to vary dramatically increases the complexity of the optimization problem. However, one could do a grid search on selection breakpoints and run CLUES2 independently for each set of breakpoints as a way to estimate when they occur. However, to obtain an accurate *P*-value for this grid search test, the log-likelihood ratio produced by CLUES2 should then be compared to a χ^2 distribution with additional degrees of freedom (an additional 1 per breakpoint). Secondly, it is also worth noting that when analyzing varying selection coefficients through time, there is the potential for overfitting if the number of different epochs is set to be too high. In particular, when considering nested models, the one with more parameters will always result in a better fit and, therefore, a higher log-likelihood ratio. Therefore, a good methodology for choosing between models should penalize additional parameters. In this manuscript, we use the AIC, which will choose a model with an additional parameter only if there is an increase in the log-likelihood ratio of at least 1. However, other approaches such as the Bayesian information criterion (Schwarz 1978) or cross validation (Yates et al. 2023) are also possible. Finally, we note that if a more complex model does not result in a substantial change in the log-likelihood ratio, an increase in the *P*-value will be observed due to the additional degrees of freedom in the corresponding χ^2 distribution.

Ancestry-Stratified Selection Analysis

The selection inference framework of CLUES2 relies on interpreting significant changes in the frequency of an allele as indicative of selection acting on that allele. However, population structure could confound the inference of selection by contributing to the change in the frequency of an allele even in the presence of selective neutrality. In particular, an observed increase in the frequency of an allele in a population could occur in the absence of selection due to either admixture from a source population where that allele is segregating at a higher frequency or due to a correlation between ancestry and sample age. To correct for this, we utilized the local ancestry labelings

available for this dataset that classify each SNP in a particular haplotype as belonging to one of four major ancestral populations that contribute to the ancestry of modern Europeans: Anatolian farmers (ANA), Caucasus hunter–gatherers (CHG), Western hunter–gatherers (WHG), and Eastern hunter–gatherers (EHG) (Pearson and Durbin 2023; Irving-Pease et al. 2024). For a given SNP, we then partition each of the ancient haplotypes into these group based on their labelings and compute the modern frequency of the allele in each ancestry group by measuring the percentage of European 1,000 Genomes haplotypes of that ancestry which have the derived allele. CLUES2 was then run for each of the four ancestries for each SNP using the number of epochs chosen by AIC in our nonstratified analysis described in the previous section. Reconstructed allele frequency trajectories and *P*-values for the selection tests are plotted in Figs. S13 to S16, Supplementary Material online. We highlight that this analysis, in addition to correcting for population structure, only utilizes ancient haplotypes and, therefore, is not affected by possible biases incurred by the miscalibration of ARG-inference methods. However, we also note that this framework is specifically designed to disentangle selection from the increase in the frequency of a neutral allele due to migration from another population and does not specifically account for adaptive introgression. This is to say that we do not model the effect of the migration of an allele from a source population resulting in the allele being brought into a new environment where it is beneficial.

Discussion

In this paper, we introduced CLUES2, a flexible, full-likelihood method that is able to utilize more of the information present in sequence data than summary statistics. CLUES2 has the ability to generate unbiased estimates of selection coefficients and also generates well-calibrated *P*-values in order to run statistical tests for selection. The methodology for our analysis of the West Eurasian dataset highlights the ability of CLUES2 to identify distinct selection coefficients in different epochs. In particular, we emphasize that different hypotheses can be tested for a given SNP without having to obtain new samples of gene trees. Furthermore, the new functionality to analyze ARGs on ancient data enables the usage of both time series data and information from linked SNPs in selection analyses.

For the ancient human DNA analysis, we observe that the one-selection coefficient model provides the best fit for the ACP2 SNP in the pan-ancestry analysis, corresponding to a classic selective sweep in progress. As increased levels of HDL have been shown to help reduce the risk of myocardial infarction (heart attack) in contemporary medical patients (Remaley et al. 2008), this result indicates that a similar benefit may have existed in ancient Eurasian populations as well. The ancestry-stratified analysis for this SNP suggests that this sweep may have been localized to WHG ancestry. The two-selection coefficient model provides a best fit for the pan-ancestry analysis of the TNXB SNP, which we see corresponds to a recent period of positive selection preceded by a period during which the derived allele was slightly disfavored. We speculate that this relatively recent selection on a gene in the HLA region, which is known to be associated with immune function (Gough and Simmonds 2007), could be the result of increasing population density and exposure to domesticated animals that occurred around this time (Page et al. 2016; Marciniak et al. 2022). The massive demographic shift induced by the introduction of agriculture likely exposed humans to new selective pressures on genes associated with immune function, explaining why we see selective sweeps on certain immunity-associated alleles beginning during this period. The ancestry-stratified analysis for this SNP suggests that this sweep may have been localized to CHG ancestry. Sparse sampling at very old time points contributes to high uncertainty in the allele frequency of the SNP for many ancestry groups.

We see that the three-selection coefficient models provide best fits for the pan-ancestry analysis of the other SNPs. For the MCM6 locus, we find a significant increase in the lactase persistence allele beginning ∼6,000 years before the present. We note that this is thousands of years after the consumption of dairy began in Europe, as evidenced by milk fat residues discovered in potsherds dating back to at least 9,000 years before the present (Evershed et al. 2022). This gap, which has previously been noted in several studies (Itan et al. 2009; Mathieson et al. 2015; Burger et al. 2020), suggests that the selective pressure for lactase persistence was not in fact the initial domestication of animals and ensuing increase in lactose consumption, an observation that has led to significant speculation as to what this pressure may have been. One hypothesis is the presence of intermittent famines during this periods caused by crop failure, during which the ability to digest alternative energy sources such as lactose would have been more advantageous (Shennan et al. 2013; Sverrisdóttir et al. 2014). Another hypothesis is that the increased population density during this time period would have caused a larger pathogen load, thus exposing individuals to more common bouts of illness (Loog et al. 2017). As the consumption of lactose by lactase non-persistent individuals can cause diarrhea and, therefore, significant fluid loss (Smith et al. 2008), it is theorized that this would result in lactase non-persistent individuals having a higher mortality when exposed to these frequent illnesses. While we do not independently verify any of these hypotheses, our finding of an increase in selective pressure for lactase persistence that significantly postdates the first appearance of sustained lactose consumption in Europe supports previous hypotheses that a separate environmental event, rather than the isolated consumption of lactose itself, is what caused the strong selection for lactase persistence. We note that the selection on lactase persistence appears to be continuing to the present day, albeit with a slightly smaller selection coefficient than when the selective pressure first began. We note that this hypothesis has

significantly better support than the hypothesis of overdominance at this locus (see the section "Selection with Dominance", Supplementary Material online). The ancestry-stratified analysis for this SNP shows that this general pattern of strong selection beginning ~6,000 years ago followed by an attenuated selection signal in the most recent time period holds for all non-Anatolian ancestries. We notice that in the WHG ancestry, selection appears to disfavor the lactase persistence allele in the most recent time period, although sparse sampling of ancient WHG haplotypes at this locus in this time period results in high uncertainty of the true allele frequency. We, therefore, caution against interpreting these results as indicating modern selection against lactase persistence in this ancestry, although our results do indicate that selection is weaker than when the selective sweep first began.

Our SLC45A2 analysis appears to capture the time in which the derived allele for rs35395 was under strong selection, which was roughly during the period of 2,500 to 5,000 years ago. We find that selection during this period was particularly strong ($s^{\mathrm{MLE}}_{2{,}500-5{,}000} = 0.07051$), but the allele appears to not be under significant selective pressure in either younger or older time periods. We note that our finding of strong selection on depigmentation in European populations beginning 5,000 years ago is in line with the results of (Wilde et al. 2014) who estimated a selection coefficient of between 2% and 10% for several skin pigmentation associated alleles. The prevailing hypothesis for the evolution of lighter skin in European populations is the resulting increased ability to synthesize vitamin D, a reaction catalyzed by solar UVB radiation, at higher latitudes where the relative concentration of solar radiation is lower (Jablonski and Chaplin 2000). Therefore, possible explanations for the exact timing of this selective sweep include the settling of environments at higher latitudes during this period and a shift away from a hunter–gatherer diet that included vitamin D-rich fish and wild game to a vitamin D-poor agriculturalist diet (Richards et al. 2003; Wilde et al. 2014). The ancestry-stratified analysis for this SNP shows that while all ancestries show evidence of very strong selection for skin depigmentation, the timing and strength of this selection does differ between ancestries. A possible explanation for this difference in the onset and strength of selection could be the differing latitudes inhabited by these ancestral groups through time and the possibly differing vitamin D content of the diets of these distinct ancestral populations.

Overall, the results of our ancient DNA analysis illustrate the importance of the introduction of agriculture to Europe and the commensurate increase in population density to changing the selective pressures associated with certain traits. With the functionality of CLUES2 to test for differing selection coefficients in different time periods, it is now possible to identify these different pressures through their effect on the historic frequency of an allele.

Despite the methodological advances of CLUES2, there are nevertheless certain limitations of this approach and areas for future research. One such limitation is the effect of possible misspecification of the data-generating process. For example, we do not consider the effect of background selection, which is known to be pervasive in the human genome and to affect the observed patterns of linked neutral mutations (Pouyet et al. 2018; Johri et al. 2021; Buffalo and Kern 2024; Cousins et al. 2024). However, we highlight the fact that the CLUES2 framework is quite flexible, allowing for possible extensions to handle model misspecifications simply by adjusting the relative weighting of the sampled trees in the importance sampling framework and without necessitating any changes to the ARG-inference methods.

Two possible sources of error when using CLUES2 on sequence data are improper ARG-sampling (either from poor mixing of the MCMC chain or inherent problems of the algorithms themselves) and using an insufficient number of importance samples in the Monte Carlo estimator. The degree to which these problems are an issue will depend on the exact dataset being analyzed. For example, in regions of the genome with high mutation to recombination rate ratios, ARG inference is easier and the posterior distribution of marginal gene trees will be so concentrated that few samples will be necessary (see Fig. 4a). However, in datasets with low mutation to recombination rate ratios and with large numbers of samples, the state space of marginal trees may be so large and the amount of information about the length of each branch may be so low that a very large number of importance samples are necessary to effectively integrate over all the uncertainty in the data. With this in mind, to be able to more efficiently utilize large numbers of importance samples, future extensions of CLUES2 may take advantage of the fact that when evaluating the likelihood at a given selection coefficient, only a small number of samples may actually contribute substantially to the sum of the Monte Carlo estimator (Equation (1)), a phenomenon known as *weight degeneracy* (Vázquez and Míguez 2017). This results in computational time being wasted in computing the forward algorithm for the samples which will not have a substantial contribution. Therefore, techniques from adaptive importance sampling or heuristics for choosing which subset of samples to use in the sum may be used, although this approach may encounter challenges when multiple selection coefficients are estimated. We note that when gene trees are correctly sampled and when a sufficient number of importance samples are used, CLUES2 vastly improves on summary statistic-based methods for inferring selection coefficients, which all seem to suffer from possible biases and significantly larger variances (see Fig. 10). Overall CLUES2 represents a significant step forward in selection inference methodology through the use of a full-likelihood model and by being able to fully utilize the information present in large ancient DNA datasets.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank the members of the Nielsen Lab at UC Berkeley for their helpful thoughts on the development of this method. In particular, we thank Diana Aguilar Gómez for her extensive testing of the software and recommendations for its input–output format and Yun Deng for his helpful discussions about HMM algorithms and associated approximations. We also thank the researchers at the Centre for GeoGenetics at the University of Copenhagen both for their feedback on CLUES2 and for hosting the authors for an extended research visit, during which much of the work for this project was done.

Data availability

The data underlying this article are available at https://erda.ku.dk/archives/917f1ac64148c3800ab7baa29402d088/published-archive.html. Publication made possible in part by support from the Berkeley Research Impact Initiative (BRII) sponsored by the UC Berkeley Library.

References

Akaike H. A new look at the statistical model identification. *IEEE Trans Auto Control*. 1974:**19**(6):716–723. https://doi.org/10.1109/tac.1974.1100705.

Allentoft ME, Sikora M, Refoyo-Martínez A, Irving-Pease EK, Fischer A, Barrie W, Ingason A, Stenderup J, Sjögren K-G, Pearson A, et al. Population genomics of post-glacial western Eurasia. *Nature*. 2024:**625**(7994):301–311. https://doi.org/10.1038/s41586-023-06865-0.

Baum LE. *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*. New York: Academic Press; 1972.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*. 2021:**220**(3):iyab229. https://doi.org/10.1093/genetics/iyab229.

Bergman J, Schrempf D, Kosiol C, Vogl C. Inference in population genetics using forward and backward, discrete and continuous time processes. *J Theor Biol*. 2018:**439**:166–180. https://doi.org/10.1016/j.jtbi.2017.12.008.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*. 2004:**74**(6):1111–1120. https://doi.org/10.1086/421051.

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.

Brandt DYC, Wei X, Deng Y, Vaughn AH, Nielsen R. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*. 2022:**221**(1):iyac044. https://doi.org/10.1093/genetics/iyac044.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*. 1995:**140**(2):783–796. https://doi.org/10.1093/genetics/140.2.783.

Buffalo V, Kern AD. A quantitative genetic model of background selection in humans. *PLoS Genet*. 2024:**20**(3):1–32. https://doi.org/10.1371/journal.pgen.1011144.

Burger J, Link V, Blöcher J, Schulz A, Sell C, Pochon Z, Diekmann Y, Žegarac A, Hofmanová Z, Winkelbach L. Low prevalence of lactase persistence in Bronze age Europe indicates ongoing strong selection over the last 3,000 years. *Curr Biol*. 2020:**30**(21):4307–4315.e13. https://doi.org/10.1016/j.cub.2020.08.033.

Chin EL, Huang L, Bouzid YY, Kirschke CP, Durbin-Johnson B, Baldiviez LM, Bonnel EL, Keim NL, Korf I, Stephensen CB, et al. Association of lactase persistence genotypes (rs4988235) and ethnicity with dairy intake in a healthy U.S. population. *Nutrients*. 2019:**11**(8):1860. https://doi.org/10.3390/nu11081860.

Cipriani V, Leung H-T, Plagnol V, Bunce C, Khan JC, Shahid H, Moore AT, Harding SP, Bishop PN, Hayward C, et al. Genome-wide association study of age-related macular degeneration identifies associated variants in the TNXB-FKBPL-NOTCH4 region of chromosome 6p21.3. *Hum Mol Genet*. 2012:**21**(18):4138–4150. https://doi.org/10.1093/hmg/dds225.

Cousins T, Tabin D, Patterson N, Reich D, Durvasula A. Accurate inference of population history in the presence of background selection. bioRxiv. https://doi.org/10.1101/2024.01.18.576291, 2024, https://www.biorxiv.org/content/early/2024/01/20/2024.01.18.576291, preprint: not peer reviewed.

Deng Y, Nielsen R, Song YS. Robust and accurate Bayesian inference of genome-wide genealogies for large samples. bioRxiv. https://doi.org/10.1101/2024.03.16.585351, 2024, https://www.biorxiv.org/content/early/2024/03/16/2024.03.16.585351, preprint: not peer reviewed.

Deshmukh N, Ganapathiraju A, Picone J. Hierarchical search for large-vocabulary conversational speech recognition: working toward a solution to the decoding problem. *IEEE Signal Process Mag*. 1999:**16**(5):84–107. https://doi.org/10.1109/79.790985.

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):233–237. https://doi.org/10.1038/ng826.

Epanechnikov VA. Non-parametric estimation of a multivariate probability density. *Theory Probab Appl*. 1969:**14**(1):153–158. https://doi.org/10.1137/1114019.

Evershed RP, Davey Smith G, Roffet-Salque M, Timpson A, Diekmann Y, Lyon MS, Cramp LJ, Casanova E, Smyth J, Whelton HL, et al. Dairying, diseases and the evolution of lactase persistence in Europe. *Nature*. 2022:**608**(7922):336–345. https://doi.org/10.1038/s41586-022-05010-7.

Fay JC, Wu C-I. Hitchhiking under positive Darwinian selection. *Genetics*. 2000:**155**(3):1405–1413. https://doi.org/10.1093/genetics/155.3.1405.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol*. 2014:**31**(5):1275–1291. https://doi.org/10.1093/molbev/msu077

Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet*. 2015:**11**(2):1–32. https://doi.org/10.1371/journal.pgen.1005004

Gough S, Simmonds M. The HLA region and autoimmune disease: associations and mechanisms of action. *Curr Genomics*. 2007:**8**(7):453–465. https://doi.org/10.2174/138920207783591690.

Grange L, Bureau J-F, Nikolayeva I, Paul R, Van Steen K, Schwikowski B, Sakuntabhai A. Filter-free exhaustive odds ratio-based genome-wide interaction approach pinpoints evidence for interaction in the HLA region in psoriasis. *BMC Genet*. 2015:**16**(1):11. https://doi.org/10.1186/s12863-015-0174-3.

Griffiths RC, Marjoram P. Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol*. 1996:**3**(4):479–502. https://doi.org/10.1089/cmb.1996.3.479.

Hejase HA, Mo Z, Campagna L, Siepel A. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *Mol Biol Evol*. 2021:**39**(1):msab332. https://doi.org/10.1093/molbev/msab332

Hudson RR. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*. 1983:**23**(2):183–201. https://doi.org/10.1016/0040-5809(83)90013-8

Illingworth CJR, Fischer A, Mustonen V. Identifying selection in the within-host evolution of influenza using viral sequence data.

PLoS Comput Biol. 2014:**10**(7):1–13. https://doi.org/10.1371/journal.pcbi.1003755

Illingworth CJR, Mustonen V. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics*. 2011:**189**(3):989–1000. https://doi.org/10.1534/genetics.111.133975

Irving-Pease EK, Refoyo-Martínez A, Barrie W, Ingason A, Pearson A, Fischer A, Sjögren K-G, Halgren AS, Macleod R, Demeter F, et al. The selection landscape and genetic legacy of ancient Eurasians. *Nature*. 2024:**625**(7994):312–320. https://doi.org/10.1038/s41586-023-06705-1

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

Jablonski NG, Chaplin G. The evolution of human skin coloration. *J Hum Evol*. 2000:**39**(1):57–106. https://doi.org/10.1006/jhev.2000.0403

Jarosz W. Efficient Monte Carlo methods for light transport in scattering media; San Diego: UC San Diego; 2008. p. 149–166. https://cs.dartmouth.edu/wjarosz/publications/dissertation/appendixA.pdf

Johri P, Riall K, Becher H, Excoffier L, Charlesworth B, Jensen JD. The impact of purifying and background selection on the inference of population history: problems and prospects. *Mol Biol Evol*. 2021:**38**(7):2986–3003. https://doi.org/10.1093/molbev/msab050

Kaplan NL, Darden T, Hudson RR. The coalescent process in models with selection. *Genetics*. 1988:**120**(3):819–829. https://doi.org/10.1093/genetics/120.3.819.

Ke W, Reed JN, Yang C, Higgason N, Rayyan L, Wählby C, Carpenter AE, Civelek M, O'Rourke EJ. Genes in human obesity loci are causal obesity genes in *C. elegans*. *PLoS Genet*. 2021:**17**(9):1–33. https://doi.org/10.1371/journal.pgen.1009736

Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. *Nat Genet*. 2019:**51**(9):1330–1338. https://doi.org/10.1038/s41588-019-0483-y

Kern AD, Schrider DR. Discoal: flexible coalescent simulations with selection. *Bioinformatics*. 2016:**32**(24):3839–3841. https://doi.org/10.1093/bioinformatics/btw556

Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*. 1969:**61**(4):893–903. https://doi.org/10.1093/genetics/61.4.893

Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, Gagnon DR, DuVall SL, Li J, Peloso GM, et al. Genetics of blood lipids among 300,000 multi-ethnic participants of the million veteran program. *Nat Genet*. 2018:**50**(11):1514–1523. https://doi.org/10.1038/s41588-018-0222-9

Le MK, Smith OS, Akbari A, Harpak A, Reich D, Narasimhan VM. 1,000 ancient genomes uncover 10,000 years of natural selection in Europe. bioRxiv. https://doi.org/10.1101/2022.08.24.505188, 2022, https://www.biorxiv.org/content/early/2022/08/26/2022.08.24.505188, preprint: not peer reviewed.

Lloyd-Jones LR, Robinson MR, Moser G, Zeng J, Beleza S, Barsh GS, Tang H, Visscher PM. Inference on the genetic basis of eye and skin color in an admixed population via Bayesian linear mixed models. *Genetics*. 2017:**206**(2):1113–1126. https://doi.org/10.1534/genetics.116.193383

Lona-Durazo F, Hernandez-Pacheco N, Fan S, Zhang T, Choi J, Kovacs MA, Loftus SK, Le P, Edwards M, Fortes-Lima CA, et al. Meta-analysis of GWA studies provides new insights on the genetic architecture of skin pigmentation in recently admixed populations. *BMC Genet*. 2019:**20**(1):59. https://doi.org/10.1186/s12863-019-0765-5

Loog L, Mirazón Lahr M, Kovacevic M, Manica A, Eriksson A, Thomas MG. Estimating mobility using sparse data: application to human genetic variation. *Proc Natl Acad Sci USA*. 2017:**114**(46):12213–12218. https://doi.org/10.1073/pnas.1703642114

Mahmoudi A, Koskela J, Kelleher J, Chan Y-B, Balding D. Bayesian inference of ancestral recombination graphs. *PLoS Comput Biol*. 2022:**18**(3):e1009960. https://doi.org/10.1371/journal.pcbi.1009960

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

Marcheco-Teruel B, Parra EJ, Fuentes-Smith E, Salas A, Buttenschøn HN, Demontis D, Torres-Español M, Marín-Padrón LC, Gómez-Cabezas EJ, Álvarez-Iglesias V, et al. Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers. *PLoS Genet*. 2014:**10**(7):1–13. https://doi.org/10.1371/journal.pgen.1004488

Marciniak S, Bergey CM, Silva AM, Hałuszko A, Furmanek M, Veselka B, Velemínský P, Vercellotti G, Wahl J, Zariņa G, et al. An integrative skeletal and paleogenomic analysis of stature variation suggests relatively reduced health for early European farmers. *Proc Natl Acad Sci USA*. 2022:**119**(15):e2106743119. https://doi.org/10.1073/pnas.2106743119

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*. 2015:**528**(7583):499–503. https://doi.org/10.1038/nature16152

Mathieson I, Terhorst J. Direct detection of natural selection in Bronze Age Britain. *Genome Res*. 2022:**32**(11–12):2057–2067. https://doi.org/10.1101/gr.276862.122

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 USA*. 2016:**113**(20):5652–5657. https://doi.org/10.1073/pnas.1514696113

Page AE, Viguier S, Dyble M, Smith D, Chaudhary N, Salali GD, Thompson J, Vinicius L, Mace R, Migliano AB. Reproductive trade-offs in extant hunter-gatherers suggest adaptive mechanism for the Neolithic expansion. *Proc Natl Acad Sci USA*. 2016:**113**(17):4694–4699. https://doi.org/10.1073/pnas.1524031113

Paris C, Servin B, Boitard S. Inference of selection from genetic time series using various parametric approximations to the Wright-Fisher model. *G3*. 2019:**9**(12):4073–4086. https://doi.org/10.1534/g3.119.400778

Pearson A, Durbin R. Local ancestry inference for complex population histories. bioRxiv. https://doi.org/10.1101/2023.03.06.529121, 2023, https://www.biorxiv.org/content/early/2023/03/07/2023.03.06.529121, preprint: not peer reviewed.

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

Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*. 2018:**7**:e36317. https://doi.org/10.7554/eLife.36317

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet*. 2014:**10**(5):e1004342. https://doi.org/10.1371/journal.pgen.1004342

Remaley AT, Amar M, Sviridov D. HDL-replacement therapy: mechanism of action, types of agents and potential clinical indications. *Expert Rev Cardiovasc Ther*. 2008:**6**(9):1203–1215. https://doi.org/10.1586/14779072.6.9.1203

Richards MP, Schulting RJ, Hedges RE. Sharp shift in diet at onset of Neolithic. *Nature*. 2003:**425**(6956):366. https://doi.org/10.1038/425366a

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 2002:**419**(6909):832–837. https://doi.org/10.1038/nature01140

Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. *Trends Genet*. 2018:**34**(4):301–312. https://doi.org/10.1016/j.tig.2017.12.005

Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978:**6**(2):461–464. https://doi.org/10.1214/aos/1176344136

Shennan S, Downey SS, Timpson A, Edinborough K, Colledge S, Kerig T, Manning K, Thomas MG. Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nat Commun*. 2013:**4**(1):2486. https://doi.org/10.1038/ncomms3486

Silverman BW. *Density estimation for statistics and data analysis*. London: Chapman and Hall; 1986.

Smith GD, Lawlor DA, Timpson NJ, Baban J, Kiessling M, Day IN, Ebrahim S. Lactase persistence-related genetic variant: population substructure and health outcomes. *Eur J Hum Genet*. 2008:**17**(3):357–367. https://doi.org/10.1038/ejhg.2008.156

Sohail MS, Louie RHY, Hong Z, Barton JP, McKay MR. Inferring epistasis from genetic time-series data. *Mol Biol Evol*. 2022:**39**(10):msac199. https://doi.org/10.1093/molbev/msac199

Sohail MS, Louie RH, McKay MR, Barton JP. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Nat Biotechnol*. 2020:**39**(4):472–479. https://doi.org/10.1038/s41587-020-0737-3

Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet*. 2019:**51**(9):1321–1329. https://doi.org/10.1038/s41588-019-0484-x

Spence JP, Zeng T, Mostafavi H, Pritchard JK. Scaling the discrete-time Wright–Fisher model to biobank-scale datasets. *Genetics*. 2023:**225**(3):iyad168. https://doi.org/10.1093/genetics/iyad168

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. https://doi.org/10.1214/14-aoas764.

Stern AJ, Wilton PR, Nielsen R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genet*. 2019:**15**(9):1–32. https://doi.org/10.1371/journal.pgen.1008384

Sverrisdóttir OÓ, Timpson A, Toombs J, Lecoeur C, Froguel P, Carretero JM, Arsuaga Ferreras JL, Götherström A, Thomas MG. Direct estimates of natural selection in Iberia indicate calcium absorption was not the only driver of lactase persistence in Europe. *Mol Biol Evol*. 2014:**31**(4):975–983. https://doi.org/10.1093/molbev/msu049

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989:**123**(3):585–595. https://doi.org/10.1093/genetics/123.3.585.

Temple SD, Waples RD, Browning SR. Modeling recent positive selection in Americans of European ancestry. bioRxiv. https://doi.org/10.1101/2023.11.13.566947, 2023, https://www.biorxiv.org/content/early/2023/11/16/2023.11.13.566947, preprint: not peer reviewed.

Terhorst J, Schlöytterer C, Song YS. Multi-locus analysis of genomic time series data from experimental evolution. *PLoS Genet*. 2015:**11**(4):1–29. https://doi.org/10.1371/journal.pgen.1005069.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015:**526**(7571):68–74. https://doi.org/10.1038/nature15393.

Tiosano D, Audi L, Climer S, Zhang W, Templeton AR, Fernández-Cancio M, Gershoni-Baruch R, Sánchez-Muro JM, Kholy ME, Hochberg Z. Latitudinal clines of the human vitamin D receptor and skin color genes. *G3*. 2016:**6**(5):1251–1266. https://doi.org/10.1534/g3.115.026773.

Torada L, Lorenzon L, Beddis A, Isildak U, Pattini L, Mathieson S, Fumagalli M. Imagene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*. 2019:**20**(S9):337. https://doi.org/10.1186/s12859-019-2927-x.

Vázquez MA, Míguez J. Importance sampling with transformed weights. *Electron Lett*. 2017:**53**(12):783–785. https://doi.org/10.1049/el.2016.3462.

Wang ET, Kodama G, Baldi P, Moyzis RK. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc Natl Acad Sci USA*. 2006:**103**(1):135–140. https://doi.org/10.1073/pnas.0509691102.

Weidinger S, Willis-Owen SAG, Kamatani Y, Baurecht H, Morar N, Liang L, Edser P, Street T, Rodriguez E, O'Regan GM, et al. A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. *Hum Mol Genet*. 2013:**22**(23):4841–4856. https://doi.org/10.1093/hmg/ddt317.

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Hollfelder N, Potekhina ID, Schier W, Thomas MG, et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc Natl Acad Sci USA*. 2014:**111**(13):4832–4837. https://doi.org/10.1073/pnas.1316513111.

Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann Math Stat*. 1938:**9**(1):60–62. https://doi.org/10.1214/aoms/1177732360.

Williamson EG, Slatkin M. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*. 1999:**152**(2):755–761. https://doi.org/10.1093/genetics/152.2.755.

Yates LA, Aandahl Z, Richards SA, Brook BW. Cross validation for model selection: a review with examples from ecology. *Ecol Monogr*. 2023:**93**(1):e1557. https://doi.org/10.1002/ecm.1557.

Ye Z, Shuai P, Zhai Y, Li F, Jiang L, Lu F, Wen F, Huang L, Zhang D, Liu X, et al. Associations of 6p21.3 region with age-related macular degeneration and polypoidal choroidal vasculopathy. *Sci Rep*. 2016:**6**(1):20914. https://doi.org/10.1038/srep20914

Zhang BC, Biddanda A, Gunnarsson ÁF, Cooper F, Palamara PF. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat Genet*. 2023:**55**(5):768–776. https://doi.org/10.1038/s41588-023-01379-x