Estimating time-varying selection coefficients from time series data of allele frequencies

Iain Mathieson[1]

November 18, 2020

1. Department of Genetics, Perelman School of Medicine, University of Pennsylvania

Abstract

Time series data of allele frequencies are a powerful resource for detecting and classifying natural and artificial selection. Ancient DNA now allows us to observe these trajectories in natural populations of long-lived species such as humans. Here, we develop a hidden Markov model to infer selection coefficients that vary over time. We show through simulations that our approach can accurately estimate both selection coefficients and the timing of changes in selection. Finally, we analyze some of the strongest signals of selection in the human genome using ancient DNA. We show that the European lactase persistence mutation was selected over the past 5,000 years with a selection coefficient of 2-2.5% in Britain, Central Europe and Iberia, but not Italy. In northern East Asia, selection at the *ADH1B* locus associated with alcohol metabolism intensified around 4,000 years ago, approximately coinciding with the introduction of rice-based agriculture. Finally, a derived allele at the *FADS* locus was selected in parallel in both Europe and East Asia, as previously hypothesized. Our approach is broadly applicable to both natural and experimental evolution data and shows how time series data can be used to resolve fine-scale details of selection.

Introduction

Time series data of allele frequencies are obtained from many sources including experimental evolution experiments and ancient DNA studies. These data are particularly useful for estimating the strength of selection and reconstructing the allele frequencies of individual alleles. This is particularly useful when timing can be informative about the basis and environmental correlates of selection.

Many methods have been developed to solve the problem of inferring selection coefficients from time series data (Bollback *et al.*, 2008; Illingworth and Mustonen, 2011; Malaspinas *et al.*, 2012; Mathieson and McVean, 2013; Nishino, 2013; Feder *et al.*, 2014; Lacerda and Seoighe, 2014; Foll *et al.*, 2015; Terhorst *et al.*, 2015; Schraiber *et al.*, 2016; Ferrer-Admetlla *et al.*, 2016; Shim *et al.*, 2016; Nené *et al.*, 2018; Paris *et al.*, 2019). One assumption common to almost all these methods is that the selection coefficient is constant throughout time. This may be appropriate in some cases, for example experimental evolution where conditions are strictly controlled, but it is less appropriate in natural populations. In particular, many of the most interesting examples of human adaptation involve adaptation to new environments, gene-culture co-evolution, or infectious diseases. Selection in these cases is likely to be time-varying, and the timing of selection is typically an important question. Inferring time-varying selection requires more data than inferring constant selection, but increasing sample sizes of ancient human DNA mean that it should now be possible to infer timings and trajectories at higher resolution.

Here, we extend the hidden Markov model of Mathieson and McVean (2013) to allow selection coefficients that change over time. A model that allowed selection coefficients to vary arbitrarily would be overfitted, so we restrict selection coefficients to a pre-specified finite number of possible values and penalize changes. By defining the model in this way we are able to compute maximum likelihood estimates of the parameters using an EM algorithm.

Methods

Wright-Fisher model

Following the notation of Mathieson and McVean (2013), we consider a Wright-Fisher population with an effective size of $2N_e$. We write f_t as the frequency of the selected allele at generation t for $t = 0 \dots T$. Suppose that the frequency trajectory is known exactly and the selection coefficient s is constant over time. Then, an approximate maximum likelihood estimator for s (Watterson, 1982) is

$$\hat{s} = \frac{f_T - f_0}{\sum_{t=0}^{T-1} f_t \left(1 - f_t\right)}. \tag{1}$$

That is, the total change in allele frequency, divided by the sum of the heterozygosity over the time the allele is observed. Now suppose that the selection coefficient at generation t is s_t, but that it takes one of K possible values $\sigma_0 \dots \sigma_K$. We assume that we know which value s_t takes at each generation and define indicator variables z_t such that $s_t = \sigma_{z_t}$. We show in the Appendix that the maximum likelihood estimator of σ_k is given by

$$\hat{\sigma_k} = \frac{\sum_{t=0}^{T-1} \mathbb{1}\left\{z_t = k\right\} \left(f_{t+1} - ft\right)}{\sum_{t=0}^{T-1} \mathbb{1}\left\{z_t = k\right\} f_t \left(1 - f_t\right)}. \tag{2}$$

This is Equation 1 with sums over generations when the selection coefficient is equal to σ_k.

Hidden Markov model - constant selection

This model was developed in Mathieson and McVean (2013), but we describe it briefly here as background for the time-varying selection mode. In practice, f_t is unknown. Instead, the data consist of samples of n_t chromosomes at each generation t (n_t can be zero), of which a_t carry the selected allele. We treat f_t as the hidden state in a hidden Markov model and (a_t, n_t) as the observations. To apply standard HMM theory, we discretize the frequency space so that $f_t \in G = \{g_1, \dots, g_D\}$, keeping the interval between grid points $\delta g = g_{i+1} - g_i$ constant. The transition probabilities $\mathbf{P}\left(f_{t+1} = g | f_t\right)$ are computed by approximating the

Wright-Fisher transition density

$$\mathbf{P}\left(f_{t+1}=g|f_t\right)=\int_{g-\delta g/2}^{g+\delta g/2}\phi\left(\frac{x-\mu_t}{\nu_t}\right)dx \tag{3}$$

where $\mu_t = f_t + sf_t\left(1-f_t\right)$ and $\nu_t = \frac{f_t(1-f_t)}{2N_e}$. The emission probabilities are binomial $a_t \sim Bin\left(n_t, f_t\right)$. We find the MLE for s by starting from an initial guess s^0 and applying the EM update rule,

$$s^{r+1}=\frac{\mathbf{E}\left[f_T\right]-\mathbf{E}\left[f_0\right]}{\sum_{t=0}^{T-1}\mathbf{E}\left[f_t\left(1-f_t\right)\right]} \tag{4}$$

with expectations over the posterior distribution of f_t computed using the forward-backward algorithm. We recalculate the forward-backward matrix and repeat until s^r converges.

Hidden Markov model - time-varying selection

In the case of time-varying selection, the hidden states are given by $\{f_t, z_t\}$ for $t = 0 \ldots T$, $f_t \in \{g_1 \ldots g_D\}$ $z_t \in \{1 \ldots K\}$ The parameters are the σ_k for $k = 1 \ldots K$ (Figure 1). The emission probabilities depend only on f_t and are the same as in the constant s model. The transition probabilities are given by

$$\mathbf{P}\left(f_{t+1}, z_{t+1}=g, j|f_t, z_t\right)=\left(c\mathbf{1}\left[j\neq z_t\right]+(1-c)\mathbf{1}\left[j=z_t\right]\right)\int_{h-\delta g/2}^{h+\delta g/2}\phi\left(\frac{x-\mu_t}{\nu_t}\right)dx \tag{5}$$

where $\mu_t = f_t + s_t f_t\left(1-f_t\right)$; $s_t = \sigma_{z_t}$; $\nu_t = \frac{f_t(1-f_t)}{N_e}$ and c is a fixed constant that gives the probability of transitioning between hidden selection states in any generation. We show in the Appendix that the EM update rule for σ_k is

$$\sigma_k^{r+1}=\frac{\sum_{t=0}^{T-1}\mathbf{E}\left[\mathbf{1}\left\{z_t=k\right\}\left(f_{t+1}-f_t\right)\right]}{\sum_{t=0}^{T-1}\mathbf{E}\left[\mathbf{1}\left\{z_t=k\right\}f_t\left(1-f_t\right)\right]}, \tag{6}$$

where now the expectations are taken over the joint posterior distribution of (f_t, z_t) calculated with the forward-backward algorithm. The forward-backward algorithm gives us the joint posterior probabilities $p_t^{g,k} = \mathbf{P}\left(f_t = g, z_t = k\right)$, which allow us to calculate the denominator and the term $\mathbf{E}\left[\mathbf{1}\left\{z_t = k\right\} f_t\right]$. To calculate the term $\mathbf{E}\left[\mathbf{1}\left\{z_t = k\right\} f_{t+1}\right]$ we also need to know the conditional posterior probabilities $p_t^{gh,kj} = \mathbf{P}\left(f_{t+1} = h, z_t = j|f_t = g, z_t = k\right)$

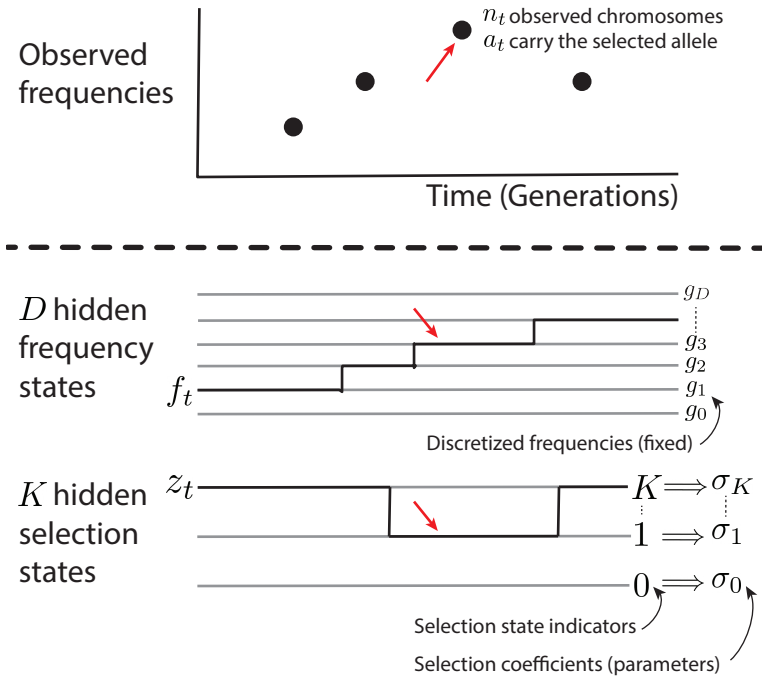

Figure 1: Schematic of the time-varying hidden Markov model. Below the dashed line are the hidden states. At the time indicated by the red arrows, we observe a_t selected alleles out of n_t total, $f_t = g_3$ and $z_t = 1$ and therefore $s_t = \sigma_{z_t} = \sigma_1$.

which can be computed from the forward and backward matrices. Then, Equation 6 can be written in terms of the discretized frequencies and posterior probabilities as

$$\sigma_k^{r+1} = \frac{\sum_{t=0}^{T-1} \sum_{g \in G} \left[p_t^{g,k} \left[\left[\sum_{h \in G} \sum_{j=1}^{K} h p_t^{gh,kj} \right] - g \right] \right]}{\sum_{t=0}^{T-1} \sum_{g \in G} \left[p_t^{g,k} g(1-g) \right]}. \tag{7}$$

In summary, the algorithm is as follows:

1. Specify the number of discrete selection coefficients, K and the per-generation probability of changing states c. Make an intial guess for the selection coefficients $\sigma_1, \ldots \sigma_K$.

2. Using the current values of $\sigma_1, \ldots \sigma_K$, the observations a_t, n_t, the binomial emission probabilities, and the transition probabilities defined in Equation 5 compute the forward and backward matrices. Use Equation 7 to update the estimates of $\sigma_1, \ldots \sigma_K$.

3. Repeat step 2 until iteration r where $\max_k |\sigma_k^r - \sigma_k^{r-1}|$ is less than some pre-defined tolerance, and stop.

Because there are DK hidden states, running time is $O(D^2K^2T)$ and space is $O(DKT)$.

Simulated data

We simulated allele frequencies under a Wright-Fisher model, with an effective population size of $N_e = 10,000$ under three different scenarios (Fig. 2A-C);

1. The selection coefficient is 0.02 for 50 generations and then -0.02 for 50 generations. Initial frequency $f_0 = 0.1$.

2. The selection coefficient is 0.02 for 100 generations, 0 for 50 generations, and then -0.02 for 50 generations. Initial frequency $f_0 = 0.1$.

3. The selection coefficient alternates between 0.02 and -0.02 every 40 generations. Initial frequency $f_0 = 0.5$.

We sampled 100 haploid individuals every 10 generations. We set initial estimates of σ_k to be ± 0.05 for $K = 2$ and $0, \pm 0.05$ for $K = 3$, a grid size of 100 (i.e. $D = 100$) and a tolerance of 0.001. We fixed the probability of transitioning between selection states c to be the inverse of the total generations observed; i.e. we expect ~ 1 selection state transition. We show the distribution of the point estimates of $\hat{\sigma}_k$, and the averaged posterior distribution of the z_t (Fig. 2D-F). Finally, we varied both the frequency and size of the samples and investigated how the performance of the estimator changed (Fig. 2G-I) in terms of:

- The root mean squared error in the estimate of the selection coefficients $\hat{\sigma}_k$.
- The posterior probability that the the inferred selection state is correct within ± 10 generations of each changepoint.
- The root mean squared error in the weighted per-generation estimate of the selection coefficient $\hat{s}_t = \sum_{g \in G} \sum_{k=1}^{K} \hat{\sigma}_k p_t^{g,k}$.

We investigated performance as we varied parameter values and specified incorrect values for fixed parameters, for example N_e, c or K.

Comparison with existing approaches

We compared our approach to CP-WFABC (Shim *et al.*, 2016)—the only existing method that is able to infer time-varying selection coefficients. Specifically, CP-WFABC uses Approximate Bayesian Computation (ABC) to fit a model with a single changepoint and two selection coefficients (i.e our scenario 1). We used the default number of simulations (1,000,000) with the best 1,000 retained, and set the prior to be the range $(-2s, 2s)$ as we tested performance for different values of s. We use the posterior distribution of the changepoint to calculate the probability of being in the wrong state, and the posterior mode as a point estimate of the selection coefficients which we compare with our maximum likelihood estimates.

Ancient DNA data

We collected published ancient DNA data from four regions of Europe chosen because they had large sample sizes and corresponding present-day data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015). We restricted to dates after the arrival of Steppe-related ancestry in each region to minimize the effects of changes in ancestry associated with that arrival (Haak *et al.*, 2015). The four regions were: Britain (GBR, 50-60°N, 5°W-2°E, <4400BP), Central Europe (CEU, 47-53°N, 8-20°E, <5000BP), Italy (TSI, 36-45°N, 7-15°E, <5000BP), Iberia (IBS, 36-44°N, 10°W-4°E, <5000BP). We identified a total of 499 samples, although not all had coverage at rs4988235 or rs174546. The samples were originally published in the following references: Allentoft *et al.* (2015); Amorim *et al.* (2018); Antonio *et al.* (2019); Fernandes *et al.* (2018); Gamba *et al.* (2014); Lipson *et al.* (2017); Martiniano *et al.* (2016, 2017); Mathieson *et al.* (2015, 2018); Mittnik *et al.* (2019); Narasimhan *et al.* (2019); Olalde *et al.* (2018, 2019); Schiffels *et al.* (2016); Valdiosera *et al.* (2018); Veeramah *et al.* (2018) and Zalloua *et al.* (2018). We also identified 255 ancient samples from East Asia (excluding Japan) from Ning *et al.* (2020); Yang *et al.* (2020) and Wang *et al.* (2020) and divided them into "North" and "South" populations at 30°N. We restricted the South population to <5000BP because only one sample was older.

Ancient DNA analysis

We used a grid of $D = 1000$, two selection states and a tolerance of 1×10^{-4}. We set N_e to grow exponentially from 10^4 to 10^6 over the past 200 years approximately as inferred by Browning and Browning (2015), though without the more rapid increase in past 10 generations. Though this estimate is for European populations, our estimator is robust to mis-specification of N_e so we assumed it was representative of late Holocene growth rates and used the same values for East Asia. Finally, we estimated the bias and uncertainty in our estimates using a parametric bootstrap: we simulated observations conditional on the inferred frequency trajectory and actual sample dates, and then reran the estimator.

Logistic regression analysis

We ran an independent analysis where we fitted the observations using logistic regression on time and ancestry components estimated using ADMIXTURE with K=3 (Alexander *et al.*, 2009). That is, the expected allele frequency of individual i, f^i is given by:

$$\log\left(\frac{f_t^i}{1-f_t^i}\right) = \beta_{P_i} t + \gamma_1 A_i + \gamma_2 B_i, \tag{8}$$

where P_i is the population to which individual i belongs and A_i and B_i are two of its ancestry component values (the third is $1 - A_i - B_i$). We estimate s by estimating the predicted change in frequency in one generation for each individual, converting it to an estimate of s based on the expected frequency change in the Wright-Fisher model (i.e. $\hat{s}^i = \frac{f_{t+1}^i - f_t^i}{f_t^i(1-f_t^i)}$) and then averaging over all individuals in each population. We estimate the standard error by assuming that the ratio of $\hat{s}$ to its standard error is the same as the ratio of β_{P_i} to its standard error. While this is not an explicit model of the evolutionary process, it does allow us to account for variation in genome-wide ancestry across individuals.

Results

Simulated data

In simulated data, we recover allele frequency trajectories, selection coefficients and the timing of changes in selection coefficients (Fig. 2). Simulations also allow us to test the robustness of the estimator to misspecification and highlight key features of its behavior. First, under scenario 1, we tested robustness to misspecification of N_e and c. These parameters must be specified in advance. However, we find that the error in the estimates is robust over one order of magnitude for N_e, and two orders of magnitude for c (Fig. S1 & S2) Thus, as long as reasonable estimates of these parameters are available, misspecification should not be a major concern. Second, we note that even for very large samples the RMSE of the selection coefficient $\hat{\sigma}_k$ and $\hat{s}_t$ do not tend to zero. This is partly due to

the stochastic effect of drift and partly due to the fact that the estimators can be biased, particularly for low initial frequencies (Fig. S3). If the initial frequency is very low, there is a relatively high chance that the allele is just by drift. For example, for an allele in a single copy, there is a probability of $\sim e^{-1} \approx 0.37$ that the allele is lost in one generation leading to a negative MLE for the selection coefficient.

As sample size increases, the RMSE of $\hat{s}_k$ decreases more reliably than that of $\hat{\sigma}_k$ (Fig. 2G-I). In other words, the estimator is better at answering the question "what is the selection coefficient in generation t?" than "what is the selection coefficient in state k?". The first question allows us to average estimates over multiple states, even if the number of states is misspecified. In fact, if there are too many or too few selection states in the HMM, then the estimator does over- or underestimate the number of transitions (Fig. S4A) but the error in $\hat{s}_t$ does not change (Fig. S4B). Therefore in our analysis of real data we focus on $\hat{s}_t$, rather than $\hat{\sigma}_k$.

In practice, the performance of the estimator depends on the data. For example, the accuracy with which we are able to detect fluctuating selection in scenario 3 (Fig. 2C) depends on the period of fluctuation (Fig. S5). Performance also depends on the sampling scheme. If we do not sample around a changepoint then we will misestimate selection coefficients around that time. Given relatively smooth trajectories, performance depends on the total number of observations—sampling ten times as many chromosomes ten times less frequently gives about the same error (Fig. 2G-I). However more uniform sampling in time would be more robust to rapidly changing trajectories. In general we recommend assessing the performance and robustness of the estimator using a parametric bootstrap approach. Run the estimator on the observed data, simulate data under the inferred model and actual pattern of observations, and investigate performance on the simulated data.

Finally, we compared the performance of our estimator to the only previously published method for detecting time-varying selection coefficients—CPWFABC (Shim *et al.*, 2016). This method uses Approximate Bayesian Computation to jointly infer a single changepoint

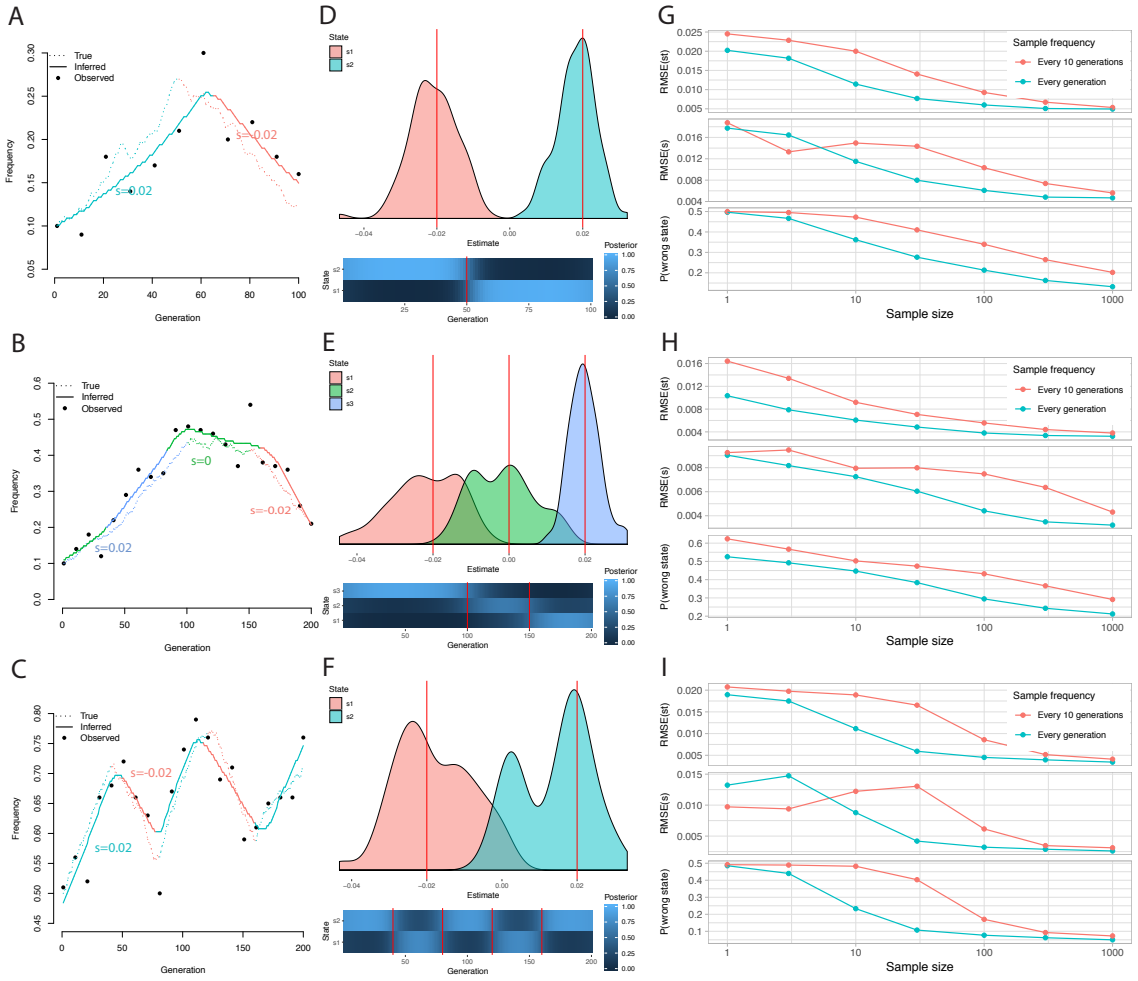

Figure 2: Performance of the estimator on simulated data. **A-C**: Simulated trajectories (dashed), observations (points), and inferred trajectories (solid). Colors indicate true and inferred selection states. **D-F**: For each of the scenarios in A-C, density plots of distribution of the estimates of the selection coefficients $\hat{\sigma}_k$ from 100 simulations. Red lines mark the true values. Lower panels show the average posterior probabilities of being in each selection state ($\mathbf{P}(z_t = k)$) in each generation. Red lines mark the true changepoints. **G-I**: For each of the scenarios in A-C, we show the RMSE error in $\hat{\sigma}_k$ in the upper panel, the RMSE error in $\hat{s}_t$ in the middle panel, and the posterior probability that z_t is wrong in ± 10 generations around each changepoint. We show estimates for sample sizes ranging from 1 to 1000, sampled either every generation or every 10 generations.

and two selection coefficients (pre- and post- changepoint). We tested the performance of this model under scenario 1 and find that our estimator outperforms it both in terms of locating the changepoint and estimating the selection coefficients (Fig. S6).

Selection at *LCT* in Europe

The SNP rs4988235 (C/T-13910) is associated with adult lactase persistence in Europeans (Enattah *et al.*, 2002) and exhibits one of the strongest signals of positive selection in the entire genome (Bersaglieri *et al.*, 2004; Grossman *et al.*, 2013). Estimates of the strength and timing of selection on the variant based on present-day data are variable and have wide confidence intervals, ranging from 0-0.2 for s) and ∼1500-65,000 years before present for the origin of the mutation (Bersaglieri *et al.*, 2004; Tishkoff *et al.*, 2007; Itan *et al.*, 2009; Peter *et al.*, 2012). Direct evidence from ancient DNA has established that the allele was rare or absent in the Neolithic and was not present at substantial frequency until the Bronze Age, starting around 5000BP (Burger *et al.*, 2007; Allentoft *et al.*, 2015; Mathieson *et al.*, 2015). In parts of Europe, for example Iberia, the derived allele did not become common until even later (Olalde *et al.*, 2019). Using ancient DNA data from across Europe, Mathieson and Mathieson (2018) estimated a selection coefficient of 0.018.

We used data from 499 ancient Europeans, divided by region, to investigate whether there were differences in the selective pressure across Europe, and whether the strength of selection varied over time (Fig. 3). We estimate that in Britain and Central Europe, the variant experienced a selection coefficient of ∼0.025, consistently for the past 4-5000 years. In Iberia, the selection coefficient was slightly lower—around 0.02. Bootstrapping suggests that the selection coefficients outside Italy might be underestimated by up to 0.005 (Fig. 3). We find no evidence that the allele was ever under selection in Italy, with an estimated selection coefficient of zero. One concern is that these differences might be due to difference in the timing of ancestry changes across Europe. We therefore fitted a logistic regression to the observations, including date and two ancestry components (inferred using

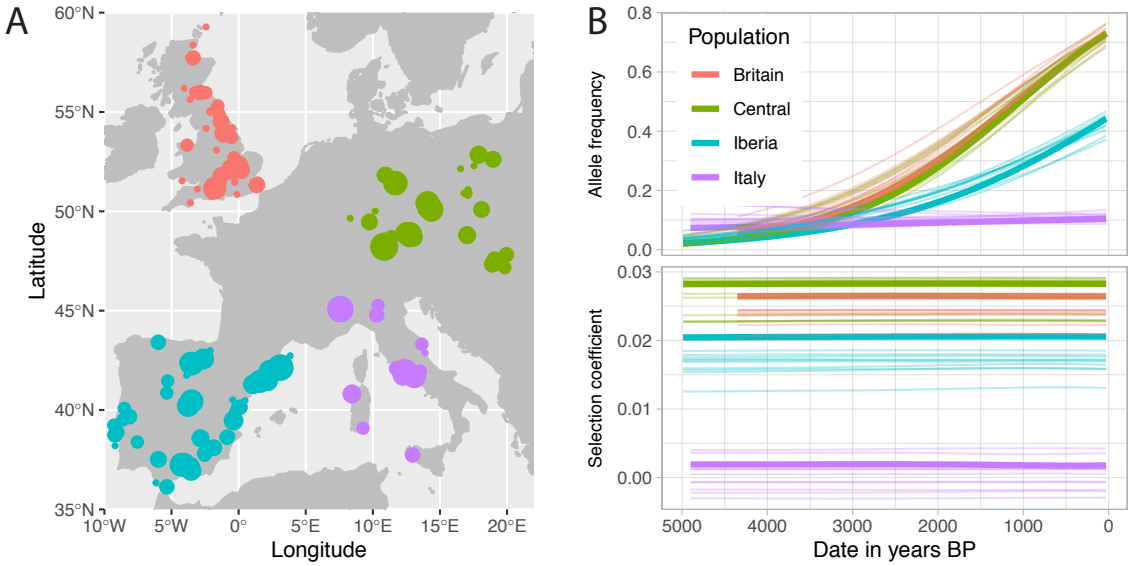

Figure 3: Selection at *LCT*. **A**: Location of 499 samples used in the analysis. The area of each circle is proportional to the sample size at each site. **B**: **Upper panel**: Solid lines indicate the inferred allele frequency trajectory for the lactase persistence allele in different parts of Europe. Faded lines indicate bootstrap replicates generated by sampling observations from this inferred frequency trajectory **Lower panel**: Inferred selection coefficient ($\hat{s}_t$) and bootstrap replicates as a function of time.

ADMIXTURE with $K = 3$). This model yields similar estimates of the selection coefficients (Fig. S7). Finally, we fitted the lattice model from Mathieson and McVean (2013) allowing migration between demes and, again, find very similar results (Fig. S8).

It is unknown whether selection on lactase persistence was dominant or additive. If we assume that the selection coefficient is constant over time, we can test the effect of different dominance parameters (Mathieson and McVean, 2013). Maximum likelihood estimates indicate complete or partial dominance, but the difference in log-likelihood is small and we cannot reject additivity (Fig. S9). Finally, it has been suggested that the allele had already reached its present-day frequency by the Middle Ages (Kruttli *et al.*, 2014) and that selection must have stopped by then. Simulations show that, given the distribution of observations, we would be unable to detect this change in selection, so this question remains unresolved (Fig. S10).

Selection at *ADH1B* in East Asia

The alcohol and aldehyde dehydrogenase genes *ADH1B* and *ALDH2* are the key components of the oxidative alcohol metabolism pathway. The derived A allele of rs1229984 in *ADH1B* increases the rate at which ethanol is oxidised to acetaldehyde and the A allele of rs671 in *ALDH2* decreases the rate at which acetaldehyde is transformed into acetic acid. The net effect of the two polymorphisms is to increase the concentration of acetaldehyde after consuming alcohol, leading to unpleasant negative effects; consequently the variants are protective against alcohol abuse (Chen *et al.*, 1999). These two variants are at high frequency in East Asia (0.8 and 0.2, respectively) compared to the rest of the world (up to 0.03 and 0.00) (1000 Genomes Project Consortium, 2015). Both variants exhibit genomic signatures of selection (Oota *et al.*, 2004; Barreiro *et al.*, 2008; Okada *et al.*, 2018). Explanations include protection against alcohol abuse and the anti-parasitic action of aldehyde (Oota *et al.*, 2004), and the variants are thought to be associated with the Neolithic development of rice farming (Peng *et al.*, 2010). Using ancient DNA from 255 ancient individuals

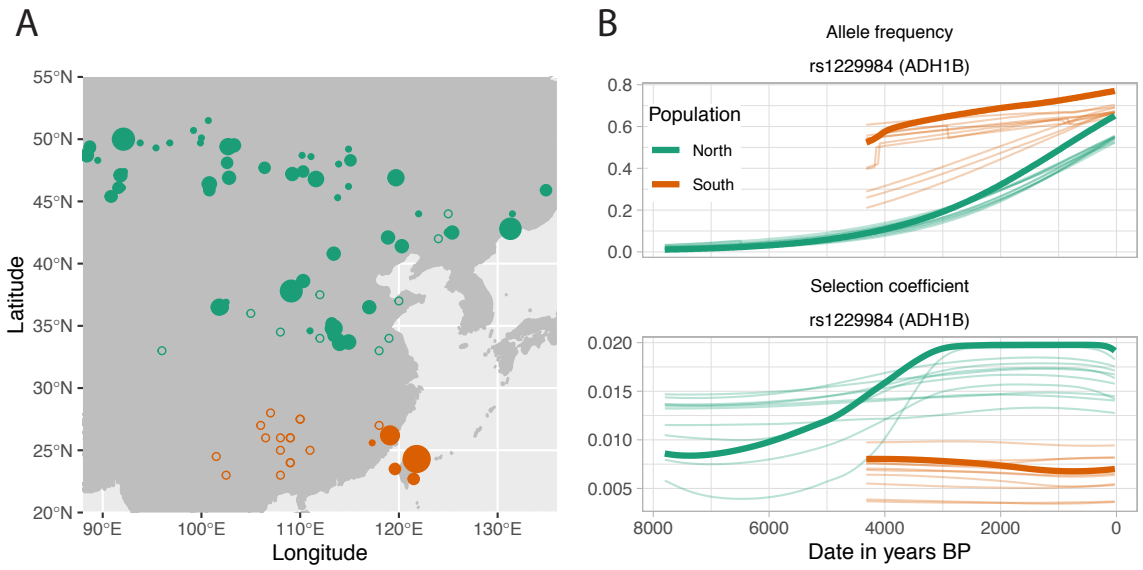

Figure 4: Selection at *ADH1B*. **A**: Location of samples used in the analysis. The area of each circle is proportional to the sample size at each site. Open circles denote locations of present-day samples. **B**: **Upper panel**: Solid lines indicate the inferred allele frequency trajectory for the derived *ADH1B* allele in North and South East Asia. Faded lines indicate bootstrap replicates generated by sampling observations from the inferred trajectory **Lower panel**: Inferred selection coefficient ($\hat{s}_t$) and bootstrap replicates as a function of time.

from East Asia (Ning *et al.*, 2020; Yang *et al.*, 2020; Wang *et al.*, 2020), and present-day allele frequencies from 1103 individuals (Peng *et al.*, 2010), we estimated the frequency and selection coefficient trajectories for *ADH1B* (Fig. 4). We estimate that by 4000 BP, the derived *ADH1B* was already common south of 30°N, but was still rare further north. Selection intensified in the north around 4000 BP with a selection coefficient of around 2%. We find consistent results if we replace the present-day population samples with the CHB and CHS 1000 Genomes populations, and when we fit the logistic regression model, correcting for $K = 3$ inferred ancestry components (Fig. S11).

Rice was domesticated in the Yangtze basin ($\approx$ 30°N) as early as 8000 BP and our results suggest that by 4000 BP, the derived *ADH1B* allele was common there. It subsequently spread north where it experienced strong selection. We did not find the derived *ALDH2* allele in any ancient individuals suggesting that it was selected in both north and south East Asia in the past few thousand years on a background of the derived *ADH1B* allele.

Selection at *FADS* in Europe and East Asia

Another signal of selection in Europe is found at the *FADS* locus. Here the derived variant has been strongly selected in the past 10,000 years and is thought to be an adaptation to an agricultural diet (Ameur *et al.*, 2012; Mathieson *et al.*, 2015; Buckley *et al.*, 2017; Ye *et al.*, 2017; Mathieson and Mathieson, 2018) In contrast to the *LCT* locus, we find that the derived allele at the *FADS* locus tagged by rs174546 follows approximately the same trajectory in each region, and has approximately the same selection coefficient (0.007-0.012), consistent with a Europe-wide estimate of 0.004-0.015 (Mathieson and Mathieson, 2018) (Fig. 5A). In East Asia, we find that the same allele has also been under recent selection, with a trajectory and selection coefficient in the north that is similar to that observed in Europe (Fig. 5B). In the south we estimate a lower frequency but stronger selection though with only one observation (out of 30) of the derived allele, this is very uncertain. In both cases, we find consistent results with the logistic regression model (Figures S12 and S13).

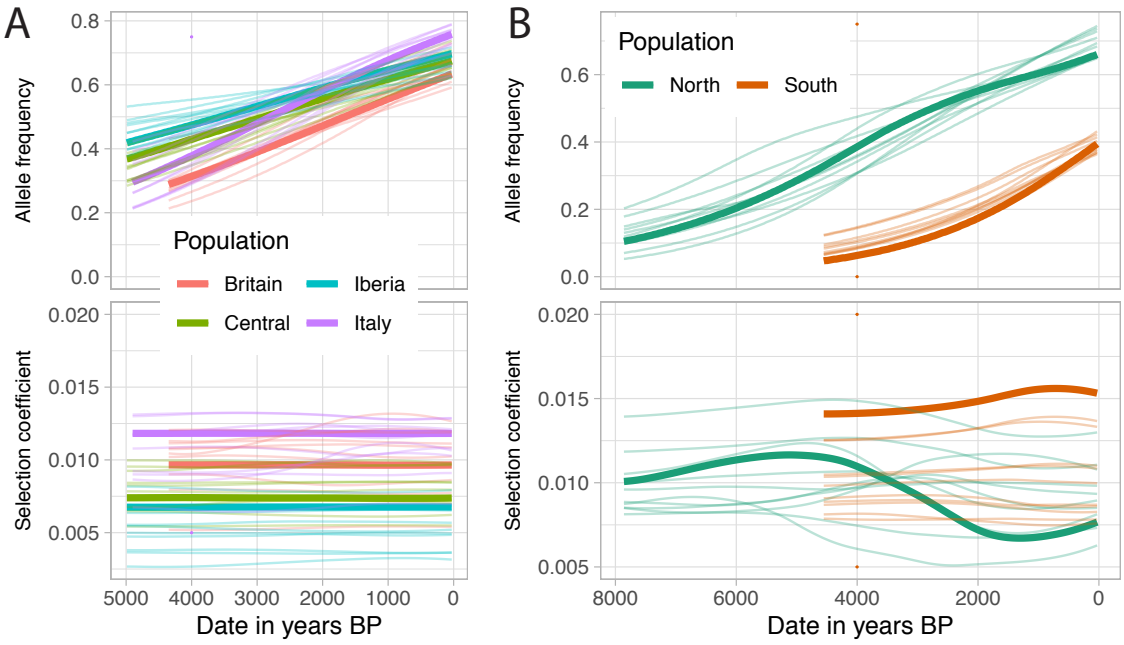

Figure 5: Inferred allele frequency trajectories and selection coefficient for the derived *FADS* allele in **A** Europe and **B** East Asia. Details are as in Figures 3 and 4. Present-day allele frequencies taken from the 1000 Genomes project populations.

Discussion

Ancient DNA is powerful tool for studying the role of natural selection in human evolution. By detecting time-varying selection, we can identify environmental changes leading to selective pressure on particular alleles. Our approach is not limited to human data, and is broadly applicable to ancient DNA, ecological or experimental evolution studies.

We find that the selection coefficient for the European lactase persistence allele was consistently around 2-2.5% in Britain, Central Europe and Iberia while the allele was not selected at all in Italy. The distribution of observations mean that we have limited power to detect changes in selection coefficient over this time period. In East Asia, our analysis of the *ADH1B* locus is consistent with selection intensifying in the North after 4000 BP, corresponding to the introduction of rice farming. However, geographic sampling and knowledge of ancestry changes is currently more limited in East Asia than in Europe, so this result does not exclude more complex trends. As previously hypothesized (Mathieson, 2020), the derived *FADS* allele was selected in both Europe and East Asia.

Genomic signatures of selection are relatively easy to detect with present-day data. Ancient DNA provides temporal information, as well as information about changes in ancestry, allowing the timing and strength of selection to be inferred. Though this does not solve the ultimate problem of identifying the environmental drivers of selection, it goes a long way to making that problem tractable, allowing hypotheses to be rejected. For example, one hypothesis about selection for lactase persistence is that it allows the uptake of vitamin D from milk rather than UV radiation, which is advantageous in the North but not South of Europe. However, our results show that selection was almost as strong in Iberia as in Northern Europe and much stronger than in Italy, making this unlikely to be the sole explanation. By allowing these inferences, our approach and others based on ancient DNA should provide much deeper insight into the nature of recent human evolution.

Acknowledgments

We thank Ziyue Gao for helpful comments on an earlier version of the manuscript. This research was funded by grants from the Alfred P. Sloan Foundation [FG-2018-10647], the Charles E. Kaufman Foundation [KA2018-98559], and NIGMS [R35GM133708]. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health or other funding sources.

Data availability

An R package is available at https://github.com/mathii/slattice/

References

1000 Genomes Project Consortium, 2015 A global reference for human genetic variation. Nature 526: 68–74.

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655–64.

Allentoft, M. E., M. Sikora, K.-G. Sjogren, S. Rasmussen, M. Rasmussen, et al., 2015 Population genomics of Bronze Age Eurasia. Nature 522: 167–172.

Ameur, A., S. Enroth, A. Johansson, G. Zaboli, W. Igl, et al., 2012 Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. Am J Hum Genet 90: 809–20.

Amorim, C. E. G., S. Vai, C. Posth, A. Modi, I. Koncz, et al., 2018 Understanding 6th-century barbarian social organization and migration through paleogenomics. Nat Commun 9: 3547.

Antonio, M. L., Z. Gao, H. M. Moots, M. Lucci, F. Candilio, et al., 2019 Ancient Rome: A genetic crossroads of Europe and the Mediterranean. Science 366: 708–714.

Barreiro, L. B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci, 2008 Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340–5.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, et al., 2004 Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74: 1111–20.

Bollback, J. P., T. L. York, and R. Nielsen, 2008 Estimation of 2Nes From Temporal Allele Frequency Data. Genetics 179: 497–502.

Browning, S. R. and B. L. Browning, 2015 Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am J Hum Genet 97: 404–18.

Buckley, M. T., F. Racimo, M. E. Allentoft, M. K. Jensen, A. Jonsson, et al., 2017 Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. Mol Biol Evol 34: 1307–1318.

Burger, J., M. Kirchner, B. Bramanti, W. Haak, and M. G. Thomas, 2007 Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proc Natl Acad Sci U S A 104: 3736–41.

Chen, C. C., R. B. Lu, Y. C. Chen, M. F. Wang, Y. C. Chang, et al., 1999 Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism. Am J Hum Genet 65: 795–807.

Enattah, N. S., T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, et al., 2002 Identification of a variant associated with adult-type hypolactasia. Nat Genet 30: 233–7.

Feder, A. F., S. Kryazhimskiy, and J. B. Plotkin, 2014 Identifying Signatures of Selection in Genetic Time Series. Genetics 196: 509–522.

Fernandes, D. M., D. Strapagiel, P. Borowka, B. Marciniak, E. Zadzinska, et al., 2018 A genomic Neolithic time transect of hunter-farmer admixture in central Poland. Sci Rep 8: 14879.

Ferrer-Admetlla, A., C. Leuenberger, J. D. Jensen, and D. Wegmann, 2016 An Approximate Markov Model for the Wright–Fisher Diffusion and Its Application to Time Series Data. Genetics 203: 831–846.

Foll, M., H. Shim, and J. D. Jensen, 2015 WFABC: a Wright–Fisher ABC-based approach

for inferring effective population sizes and selection coefficients from time-sampled data. Molecular Ecology Resources 15: 87–98.

Gamba, C., E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes, et al., 2014 Genome flux and stasis in a five millennium transect of European prehistory. Nat Commun 5: 5257.

Grossman, S. R., K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, et al., 2013 Identifying recent adaptations in large-scale genomic data. Cell 152: 703–13.

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, et al., 2015 Massive migration from the steppe is a source for Indo-European languages in Europe. Nature 522: 207–11.

Illingworth, C. J. R. and V. Mustonen, 2011 Distinguishing Driver and Passenger Mutations in an Evolutionary History Categorized by Interference. Genetics 189: 989–1000.

Itan, Y., A. Powell, M. A. Beaumont, J. Burger, and M. G. Thomas, 2009 The origins of lactase persistence in Europe. PLoS Comput Biol 5: e1000491.

Kruttli, A., A. Bouwman, G. Akgul, P. Della Casa, F. Ruhli, et al., 2014 Ancient DNA analysis reveals high frequency of European lactase persistence allele (T-13910) in medieval central europe. PLoS One 9: e86251.

Lacerda, M. and C. Seoighe, 2014 Population Genetics Inference for Longitudinally-Sampled Mutants Under Strong Selection. Genetics 198: 1237–1250.

Lipson, M., A. Szecsenyi-Nagy, S. Mallick, A. Posa, B. Stegmar, et al., 2017 Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature 551: 368–372.

Malaspinas, A.-S., O. Malaspinas, S. N. Evans, and M. Slatkin, 2012 Estimating Allele Age and Selection Coefficient from Time-Serial Data. Genetics 192: 599–607.

Martiniano, R., A. Caffell, M. Holst, K. Hunter-Mann, J. Montgomery, et al., 2016 Genomic signals of migration and continuity in Britain before the Anglo-Saxons. Nature communications 7: 10326.

Martiniano, R., L. M. Cassidy, R. O'Maolduin, R. McLaughlin, N. M. Silva, et al., 2017 The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. PLoS Genet 13: e1006852.

Mathieson, I., 2020 Limited Evidence for Selection at the FADS Locus in Native American Populations. Molecular Biology and Evolution 37: 2029–2033.

Mathieson, I., S. Alpaslan-Roodenberg, C. Posth, A. Szecsenyi-Nagy, N. Rohland, et al., 2018 The genomic history of southeastern Europe. Nature 555: 197–203.

Mathieson, I., I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, et al., 2015 Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528: 499–503.

Mathieson, I. and G. McVean, 2013 Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics 193: 973–84.

Mathieson, S. and I. Mathieson, 2018 FADS1 and the Timing of Human Adaptation to Agriculture. Mol Biol Evol 35: 2957–2970.

Mittnik, A., K. Massy, C. Knipper, F. Wittenborn, R. Friedrich, et al., 2019 Kinship-based social inequality in Bronze Age Europe. Science 366: 731–734.

Narasimhan, V. M., N. Patterson, P. Moorjani, N. Rohland, R. Bernardos, et al., 2019 The formation of human populations in South and Central Asia. Science 365: eaat7487.

Nené, N. R., A. S. Dunham, and C. J. R. Illingworth, 2018 Inferring Fitness Effects from Time-Resolved Sequence Data with a Delay-Deterministic Model. Genetics 209: 255–264.

Ning, C., T. Li, K. Wang, F. Zhang, T. Li, et al., 2020 Ancient genomes from northern China suggest links between subsistence changes and human migration. Nat Commun 11: 2700.

Nishino, J., 2013 Detecting Selection Using Time-Series Data of Allele Frequencies with Multiple Independent Reference Loci. G3: Genes, Genomes, Genetics 3: 2151–2161.

Okada, Y., Y. Momozawa, S. Sakaue, M. Kanai, K. Ishigaki, et al., 2018 Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. Nat Commun 9: 1631.

Olalde, I., S. Brace, M. E. Allentoft, I. Armit, K. Kristiansen, et al., 2018 The Beaker phenomenon and the genomic transformation of northwest Europe. Nature 555: 190–196.

Olalde, I., S. Mallick, N. Patterson, N. Rohland, V. Villalba-Mouco, et al., 2019 The genomic history of the Iberian Peninsula over the past 8000 years. Science 363: 1230–1234.

Oota, H., A. J. Pakstis, B. Bonne-Tamir, D. Goldman, E. Grigorenko, et al., 2004 The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. Ann Hum Genet 68: 93–109.

Paris, C., B. Servin, and S. Boitard, 2019 Inference of Selection from Genetic Time Series Using Various Parametric Approximations to the Wright-Fisher Model. G3: Genes, Genomes, Genetics 9: 4073–4086.

Peng, Y., H. Shi, X. B. Qi, C. J. Xiao, H. Zhong, et al., 2010 The ADH1B Arg47His polymorphism in east Asian populations and expansion of rice domestication in history. BMC Evol Biol 10: 15.

Peter, B. M., E. Huerta-Sanchez, and R. Nielsen, 2012 Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet 8: e1003011.

Schiffels, S., W. Haak, P. Paajanen, B. Llamas, E. Popescu, et al., 2016 Iron age and Anglo-Saxon genomes from East England reveal British migration history. Nature communications 7: 10408.

Schraiber, J. G., S. N. Evans, and M. Slatkin, 2016 Bayesian Inference of Natural Selection from Allele Frequency Time Series. Genetics 203: 493–511.

Shim, H., S. Laurent, S. Matuszewski, M. Foll, and J. D. Jensen, 2016 Detecting and Quantifying Changing Selection Intensities from Time-Sampled Polymorphism Data. G3 (Bethesda) 6: 893–904.

Terhorst, J., C. Schlötterer, and Y. S. Song, 2015, 04)Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution. PLOS Genetics 11: 1–29.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, et al., 2007 Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31–40.

Valdiosera, C., T. Günther, J. C. Vera-Rodríguez, I. Ureña, E. Iriarte, et al., 2018 Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. Proceedings of the National Academy of Sciences 115: 3428–3433.

Veeramah, K. R., A. Rott, M. Groß, L. van Dorp, S. López, et al., 2018 Population genomic analysis of elongated skulls reveals extensive female-biased immigration in Early Medieval Bavaria. Proceedings of the National Academy of Sciences 115: 3494–3499.

Wang, C.-C., H.-Y. Yeh, A. N. Popov, H.-Q. Zhang, H. Matsumura, et al., 2020 The Genomic Formation of Human Populations in East Asia. bioRxiv.

458 Watterson, G. A., 1982 Testing selection at a single locus. Biometrics 38: 323–331.

459 Yang, M. A., X. Fan, B. Sun, C. Chen, J. Lang, et al., 2020 Ancient DNA indicates human
460 population shifts and admixture in northern and southern China. Science.

461 Ye, K., F. Gao, D. Wang, O. Bar-Yosef, and A. Keinan, 2017 Dietary adaptation of FADS
462 genes in Europe varied across time and geography. Nat Ecol Evol 1: 167.

463 Zalloua, P., C. J. Collins, A. Gosling, S. A. Biagini, B. Costa, et al., 2018 Ancient DNA of
464 Phoenician remains indicates discontinuity in the settlement history of Ibiza. Sci Rep 8:
465 17567.

Appendix

This derivations follow very closely those for the constant selection case in Mathieson and McVean (2013). Suppose the allele frequency f_t in generation t is known exactly. The selection coefficient in generation t is $s_t = \sigma_k \mathbb{1}\{Z_t = k\}$ where z_t is known. Then, conditional on f_t, the distribution of f_{t+1} is binomial with size N_e and probability $f_t + s_t f_t(1 - f_t)$. Thus, log-likelihood of the selection coefficients $\sigma_1 \dots \sigma_K$ is given by:

$$\ell(\sigma_1, \dots, \sigma_K) = 2N_e \sum_{t=1}^{T} \{f_t \log(1 + s_t) - \log(1 + s_t f_{t-1}\}. \tag{9}$$

But, since $s_t = \sigma_k \mathbb{1}\{Z_t = k\}$, the log-likelihoods for each σ_k do not depend on each other so we can write

$$\ell(\sigma_k) = 2N_e \sum_{t=1}^{T} \{f_t \log(1 + \sigma_k \mathbb{1}\{Z_t = k\}) - \log(1 + \sigma_k \mathbb{1}\{Z_t = k\} f_{t-1})\}. \tag{10}$$

Differentiating w.r.t. σ_k and setting equal to zero gives.

$$\sum_{t=1}^{T} \left\{ \frac{f_{t-1}(1 + \hat{\sigma}_k)\mathbb{1}\{Z_t = k\}}{1 + f_{t-1}\hat{\sigma}_k \mathbb{1}\{Z_t = k\}} \right\} - \sum_{t=1}^{T} f_t \mathbb{1}\{Z_t = k\} = 0. \tag{11}$$

Expanding the fraction to first order in σ_k gives

$$\sum_{t=1}^{T} \mathbb{1}\{Z_t = k\} \left\{ (f_{t-1}(1 + \hat{\sigma}_k))(1 - f_{t-1}\hat{\sigma}_k) - f_t + O(\hat{\sigma}_k^2) \right\} = 0. \tag{12}$$

$$\hat{\sigma}_k \sum_{t=1}^{T} \{(f_{t-1}(1 - f_{t-1}\} - \sum_{t=1}^{T} \{(f_t - f_{t-1}\} + O(\hat{\sigma}_k^2) = 0, \tag{13}$$

which yields the result in Equation 2. Another way to see this is that in Equation 10, we could remove the indicator functions and write the sum over $t : Z_t = k$, rather than $t = 1 \dots T$ leading to an equivalent form of Equation 2. For the EM update step we maximize the expectation over $\{f_t, z_t\}$ of the likelihood (Equation 10). Taking expectations, differentiating and setting equal to zero we obtain, by the same argument above, the result of Equation 6.

Supplementary Figures

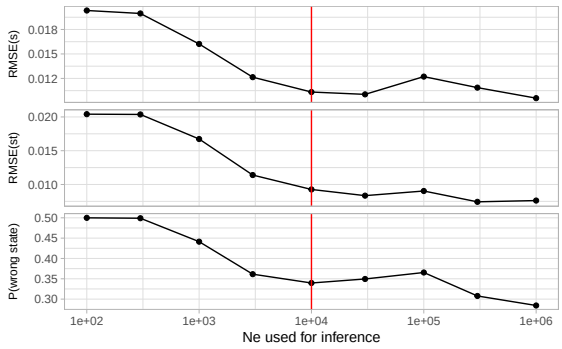

Figure S1: Errors in scenario 1 (defined as in Fig. 2G) when N_e is mis-specified. True $N_e = 10,000$, and we sample 100 chromosomes every 10 generations.

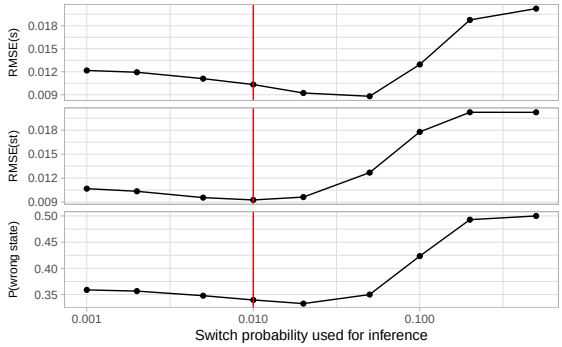

Figure S2: Errors in scenario 1 (defined as in Fig. 2G) when c is mis-specified. True $c = 0.01$, and we sample 100 chromosomes every 10 generations.

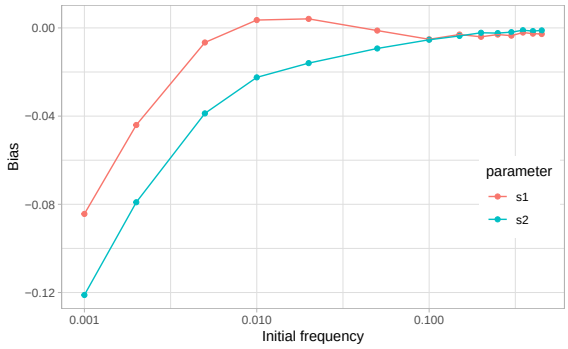

Figure S3: Bias in the estimate of selection coefficients $\hat{\sigma}_k$ in scenario 1 as a function of initial allele frequency. Simulations as in Fig. 2D.

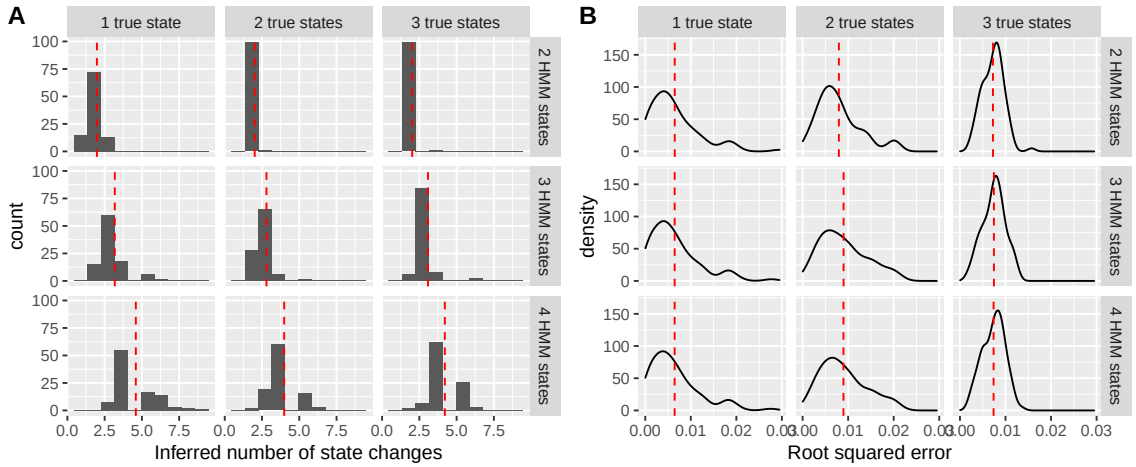

Figure S4: Performance of the estimator when the number of selection states is misspecified. **A**: distribution of the number of inferred state changes (in the sense that the most likely state changes), for different numbers of true model states. Histograms show the distribution of inferred state changes from 100 replicates, and dashed red lines show the mean. For 1 true state we simulate $s = 0.02$ for 50 generations, for 2 states we simulate $s = 0.02$ and 0 for 50 generations each, and for 3 states we simulate $s = 0.02$, 0 and -0.02 for 50 generations each. **B**: With the same simulations as part A, we show the distribution of RMSE of $\hat{s}_t$ for different numbers of model states. Dashed red lines show the mean.

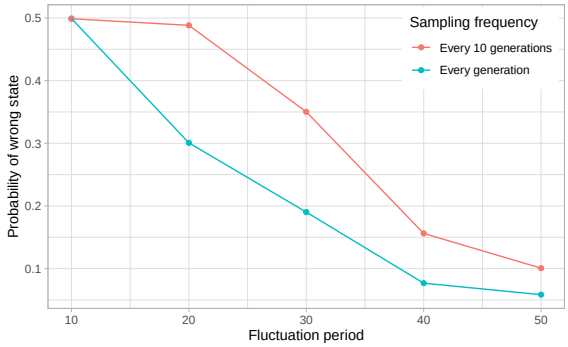

Figure S5: Performance of the estimator for scenario 3 (Fig. 2C) when the period of fluctuation varies. We show the probability that we estimate that we are in the wrong state. Observations are 100 chromosomes either every generation or every 10 generations.

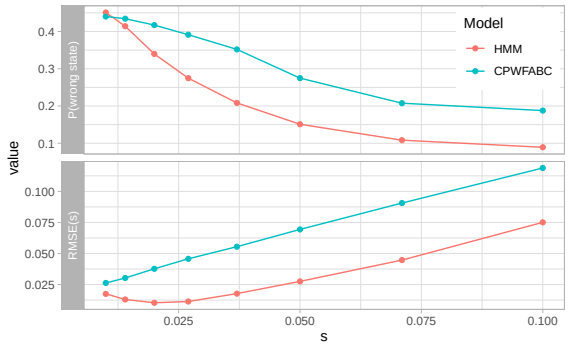

Figure S6: Performance comparison with CP-WFABC. We show the probability of being in the wrong state ± 10 generations around the true changepoint, and the average error in the estimated selection coefficient (i.e. $\hat{\sigma}_k$ for our HMM and the CP-WFABC posterior mode).

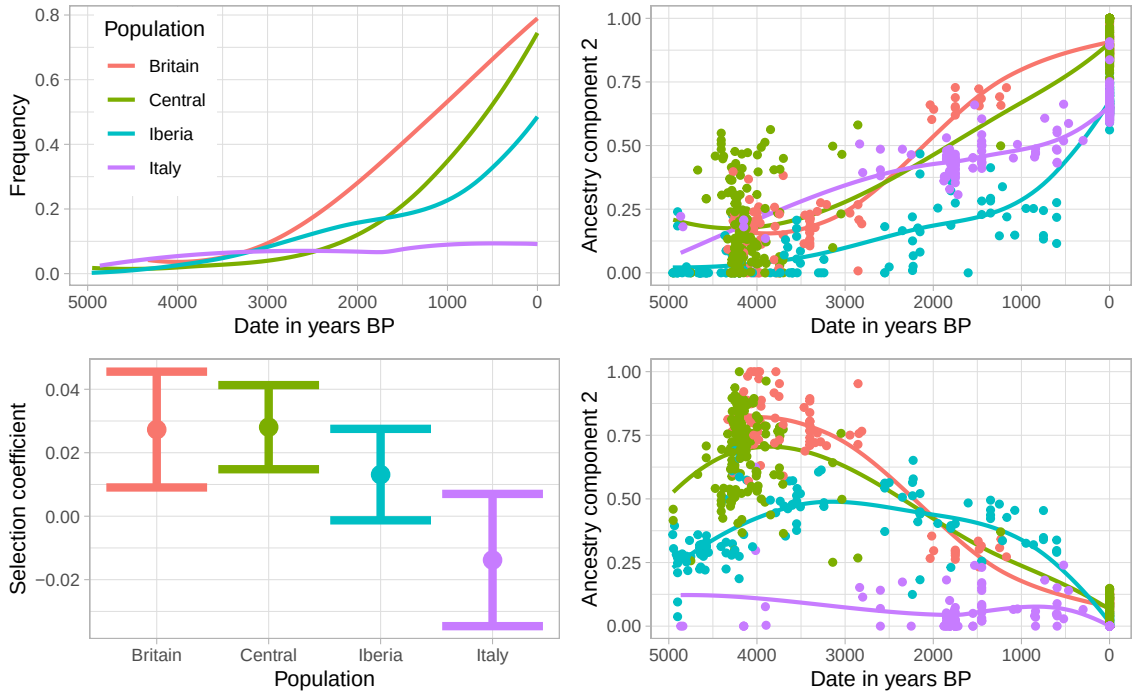

Figure S7: Results of fitting a logistic regression to the observations of the derived *LCT* allele, as a function of date and ancestry (inferred using ADMIXTURE with $K = 3$, and converting the effect size for date to an estimate of the selection coefficient (Methods). **Top left**: LOESS smoothed fitted allele frequency trajectories in each region. **Top left**: Estimated selection coefficients and 95% confidence intervals in each region **Right panels**: Ancestry components for each individual, with smoothed LOESS fit lines.

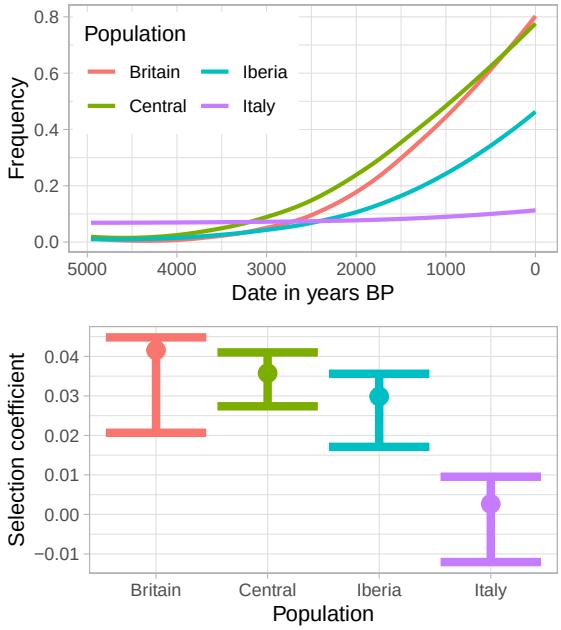

Figure S8: Results of fitting the 2×2 lattice model of (Mathieson and McVean, 2013) to the data, allowing migration between Britain and Iberia, Britain and Central, Central and Italy, and Italy and Iberia. **Upper panel**: Inferred allele frequency trajectories in each region. **Lower panel**: Estimated selection coefficients and approximate 95% confidence intervals in each region.

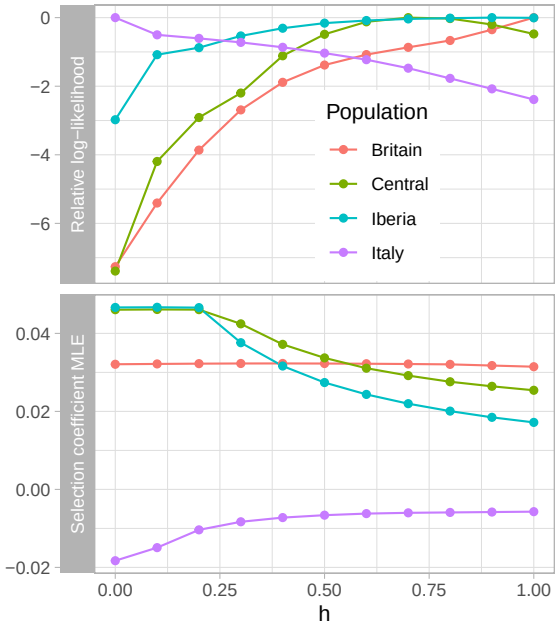

Figure S9: Results of fitting the single population model of Mathieson and McVean (2013) and allowing the dominance parameter h to vary. **Upper panel**: Log-likelihood (relative to the maximum) as a function of the dominance parameter h. **Lower panel**: Maximum likelihood estimate of s as a function of the dominance parameter h.

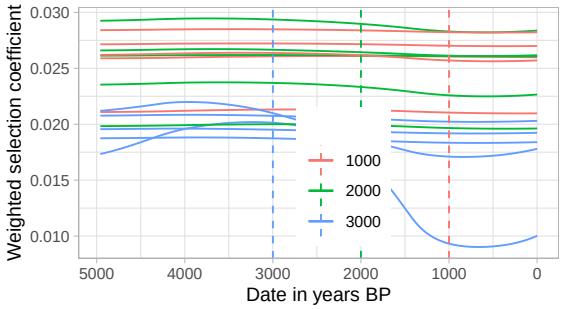

Figure S10: Testing whether we can detect the end of selection on *LCT*. Keeping the existing sampling points, we made the inferred allele frequency trajectory 1000, 2000 or 3000 years shorter, keeping the same total increase in frequency and inserting a 500,1000 or 1500 year period of constant frequency until the present. We then simulated observations keeping the observed distribution, and reran the estimator. We show 5 replicate simulations for each estimator.

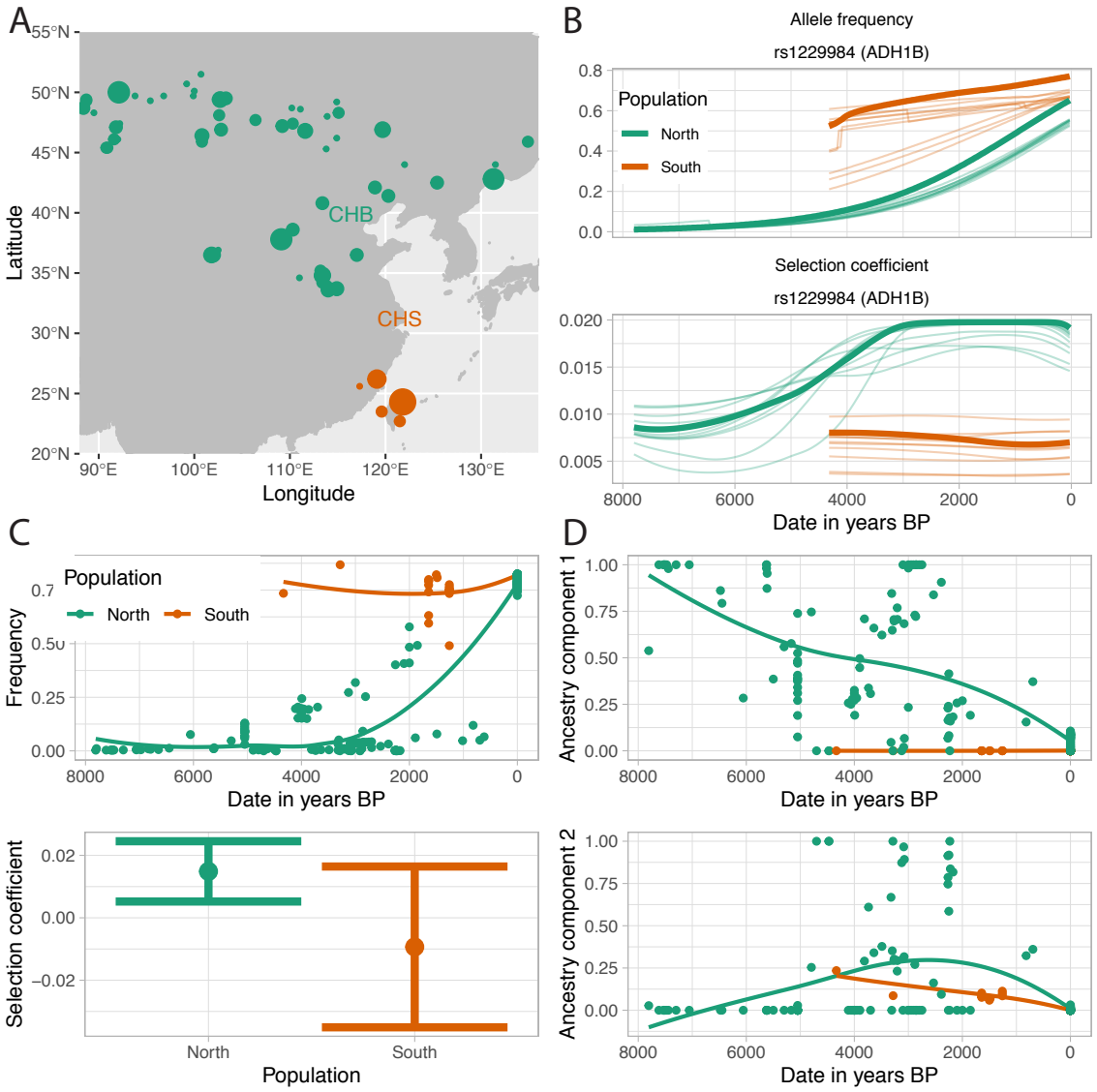

Figure S11: **A**: Location of present-day CHB and CHS populations from 1000 Genomes. **B**: Inferred frequency trajectories and selection coefficients for the derived *ADH1B* allele using present-day 1000 Genomes population frequencies (CHB/CHS). **C**: Inferred allele frequency trajectory and (constant) selection coefficient for the logistic regression model. Points show the fitted values for each ancient individuals and lines show a LOESS fit. **D**: Two ancestry components inferred using ADMIXTURE. Points show the fitted values and lines show a LOESS fit.

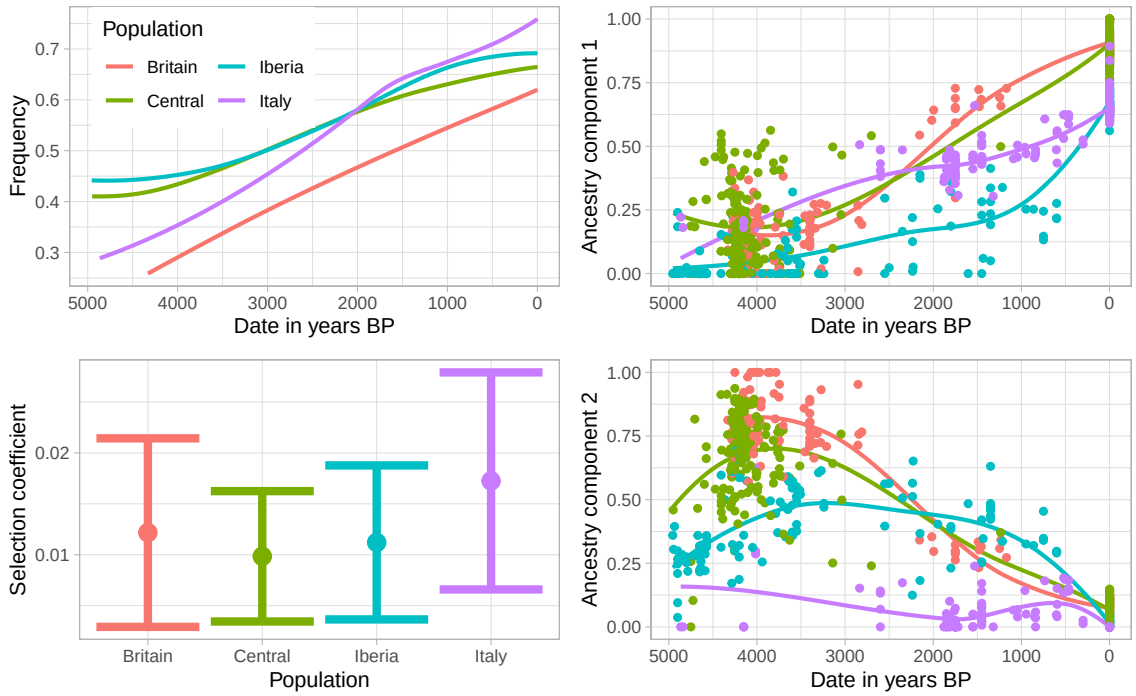

Figure S12: Results of fitting a logistic regression to the observations of the derived *FADS* allele in Europe, as a function of date and ancestry (inferred using ADMIXTURE with $K = 3$), and converting the effect size for date to an estimate of the selection coefficient (Methods). **Upper left**: Fitted allele frequency trajectories in each region. **Lower left**: Estimated selection coefficients and 95% confidence intervals in each region **Right panels**: Ancestry components for each individual (identical to Figure S7), with region-specific smoothed loess fit lines.

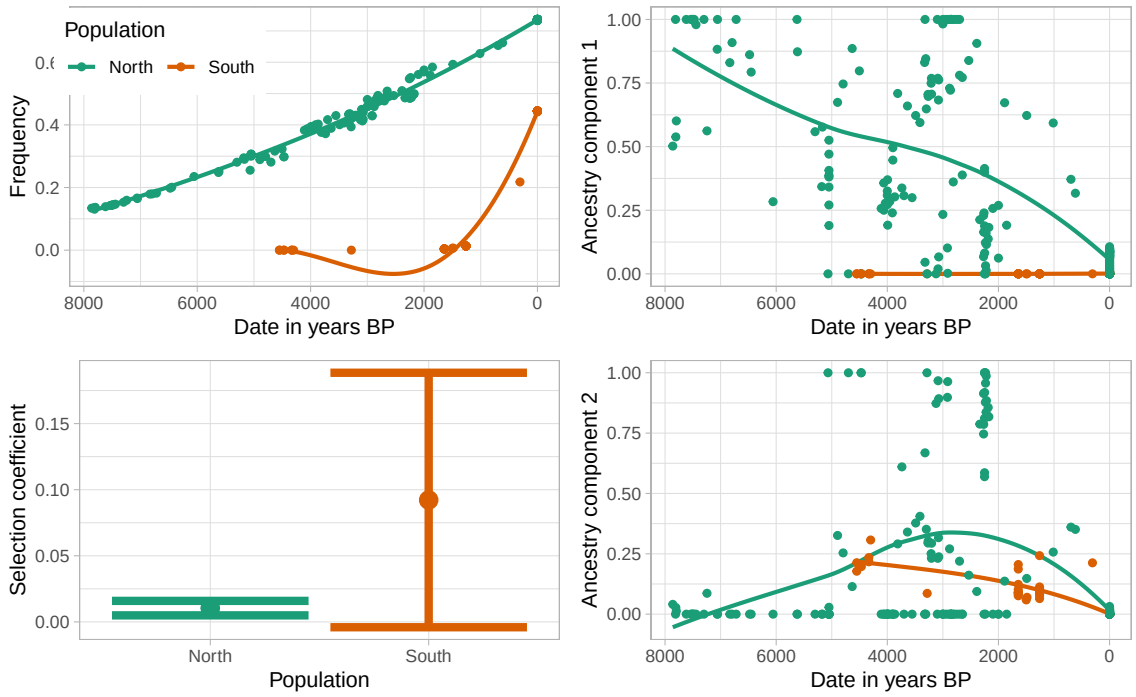

Figure S13: Results of fitting a logistic regression to the observations of the derived *FADS* allele in East Asia, as a function of date and ancestry (inferred using ADMIXTURE with $K = 3$), and converting the effect size for date to an estimate of the selection coefficient (Methods). **Upper left**: Fitted allele frequency trajectories in each region. **Lower left**: Estimated selection coefficients and 95% confidence intervals in each region (0.004-0.015 and -0.01-0.18 in North and South, respectively). **Right panels**: Ancestry components for each individual (identical to Figure S11), with region-specific smoothed LOESS fit lines.