bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Robust detection of natural selection using a probabilistic model of tree imbalance

Enes Dilber and Jonathan Terhorst*
Department of Statistics, University of Michigan

May 12, 2021

Abstract

Neutrality tests such as Tajima's D (Tajima, 1989) and Fay and Wu's H (Fay and Wu, 2000) are standard implements in the population genetics toolbox. One of their most common uses is to scan the genome for signals of natural selection. However, it is well understood that deviance measures like D and H are confounded by other evolutionary forces—in particular, population expansion—that may be unrelated to selection. Because they are not model-based, it is not clear how to deconfound these statistics in a principled way.

In this paper we derive new likelihood-based methods for detecting natural selection which are robust to confounding by fluctuations in effective population size. At the core of our method is a novel probabilistic model of tree imbalance, which generalizes Kingman's coalescent to allow certain aberrant tree topologies to arise more frequently than is expected under neutrality. We derive a frequency spectrum-based estimator which can be used in place of D, and also extend to the case where genealogies are first estimated. We benchmark our methods on real and simulated data, and provide an open source software implementation.

1 Introduction

Understanding how species to adapt to their surroundings has been a defining challenge in biology for several centuries. One of the primary drivers of adaptation is, of course, natural selection. Recently, as genomic data has become much easier to obtain, significant efforts have been made to study

*Corresponding author: jonth@umich.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

natural selection using patterns of population genetic variation. In addition to advancing our general knowledge of evolution, this research has the potential to improve health and reduce disease by pinpointing the molecular basis for certain complex, adaptive phenotypes.

Because natural selection exerts a strong influence on the trajectory (frequency over time) of a selected allele, the ideal data for studying selection are time series of allele frequencies observed across many generations. Unfortunately, such data are rare except in laboratory settings. In order to study selection in natural populations, research has focused on devising methods for inferring selection from contemporaneous samples of polymorphism data. This is a challenging problem, because we have to make inferences about complex selection mechanisms using only a “snapshot” of genetic variation obtained at a single point in time. Theoretical models are essential in order to decipher these complicated signals in a principled way.

One way to reason about signals of natural selection is by considering its effect on genealogies. Relative to a neutral baseline, natural selection induces certain genealogical distortions. For example, a positively-selected variant sweeping towards fixation induces unbalanced, “star-like” genealogies, resulting in excesses of linkage disequilibrium and low- and high-frequency variants in the vicinity of the selected allele (Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000; Kim and Nielsen, 2004). Another form, balancing selection, produces genealogies which outwardly resemble those found in a structured population (Kaplan, Darden, and Hudson, 1988). These distortions are then manifested in terms of altered patterns of genetic variation. By fitting a statistical model of this process, we can learn about natural selection from polymorphism data.

1.1 Our contribution

In this article, we derive new procedures for detecting natural selection in genetic variation data. Our approach is based on a probabilistic model of genealogical imbalance which is designed to capture certain hallmark signals of selection described above. It generalizes Kingman’s ubiquitous coalescent process (Kingman, 1982a; Kingman, 1982b), and builds on earlier attempts in phylogenetics to model the process of speciation (Aldous, 1996; Blum and François, 2006). Although more principled and correct models of the coalescent process under selection have been studied previously (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997), owing to their complexity, they are not widely used for inference. As we will see, ours is a simple approximation which retains much of tractability of neutral coalescent; the

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

resulting estimators are fast, model-based, and easy to understand and implement. An important feature of our method it explicitly models variation in effective population size, leading to a "demographically corrected" neutrality test that has demonstrable advantages when population size indeed varies over time. Finally, because our method is based on a generative model of tree formation, it can be extended with little effort to cases where gene trees or ancestral recombination graphs have already been inferred, as is becoming increasingly common in population genetics (Kelleher et al., 2019a; Speidel et al., 2019).

1.2 Related work

We lack space to survey the full panoply of methods that have been developed to study natural selection using genomic data; see recent reviews by Vitti, Grossman, and Sabeti (2013) and Stern and Nielsen (2019). We focus here on two classes of methods for detecting natural selection which are most closely related to our proposed approach.

The first class is *frequency spectrum-based methods*, which operate on the principle that natural selection distorts equilibrium allele frequencies relative to what is observed under neutrality. The most widely used frequency spectrum-based statistic is Tajima's D (Tajima, 1989):

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\hat{s}}, \tag{1}$$

where $\hat{\theta}_\pi$ and $\hat{\theta}_W$ are, respectively, Tajima's and Watterson's estimators of the population-scaled mutation rate θ, and $\hat{s}$ is an estimate of the standard deviation of their difference. Both estimators are unbiased for θ under neutrality, but have different biases for non-neutral evolution, such that $\mathbb{E}D \neq 0$ when examining allele frequencies obtained from a region that is under selection. Other related statistics include Fu and Li's D (Fu and Li, 1993), and Fay and Wu's H (Fay and Wu, 2000). A unifying interpretation of the various frequency spectrum-based statistics was given by Achaz (2009) who showed that each can be written as a certain weighted sum of entries of the SFS.

As suggested by (1), a common feature shared by all of the abovementioned tests is that they are based on measures of deviance. That is, under neutrality each test statistic has zero mean, and larger magnitudes of the statistic suggest larger deviations from neutrality. However, beyond this general feature, interpretation of these measures can be subtle. For

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

example, Tajima's D is sensitive to deviations at all locations of the frequency spectrum, whereas Fay and Wu's H only has power to detect a large excess of high frequency variants (Achaz, 2009). Negative values of D might indicate either directional selection or population growth, while positive D can alternatively indicate either balancing selection or population structure (Ferretti et al., 2017). More generally, deviance statistics based on the SFS are confounded by other evolutionary forces, in particular fluctuating historical effective population size, and there is not an obvious way to compensate for this[1]. Finally, because they operate using only marginal allele frequency information, these methods do not incorporate haplotype information or patterns of allele sharing, which can be a valuable auxiliary signal of natural selection.

A second group of methods for detecting selection can be described as *haplotype-based* methods. These are designed to exploit characteristic signatures of linkage disequilibrium that are deposited in the genome in the wake of a selective event (Maynard Smith and Haigh, 1974; Kaplan, Hudson, and Langley, 1989). Among the best-known of this class of methods are the so-called extended haplotype homozygosity (EHH) score (Sabeti et al., 2006), the integrated haplotype score (iHS; Voight et al., 2006), and the singleton density score (SDS; Field et al., 2016). Each of these scores is derived via population genetic and/or genealogical arguments about how variation is altered in the vicinity of a selected variant. For example, SDS is designed to detect regions of the genome where the terminal branches of the underlying genealogy are shorter than usual, as is expected under recent positive selection. However, although each of these statistics has been shown to work well in certain settings, ultimately these methods are heuristic, and not based on a concrete evolutionary model.

Given the profusion of *ad hoc* methods that have been proposed for detecting natural selection, it is natural to wonder why likelihood-based methods are not more common. The advantages of likelihood-based testing and estimation are well known (Neyman and Pearson, 1933; Lehmann and Casella, 2006). However, likelihood-based methods in population genetics are, in general, difficult: computing the likelihood of a sample of genomes, even under a simple neutral model, requires integrating over all of the possible ancestry scenarios that could have generated a given data set, a massive computational undertaking (Stern and Nielsen, 2019). Nevertheless, there

[1] One standard practice is to subtract the genomewide mean of the test statistic from local estimates. But this assumes that the bulk of the genome is evolving neutrally, and recent work has questioned the validity of this assumption (McVicker et al., 2009; Cai et al., 2009; Lohmueller et al., 2011).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

has been some recent progress. Berg and Coop (2015) studied an approximate likelihood model for selection at a single locus, and very recently, a noteworthy contribution was made by Stern, Wilton, and Nielsen (2019), who propose an approximate full-likelihood method for inferring natural selection using recombining sequence data. Building on earlier work (Rasmussen et al., 2014), their method (approximately) integrates over the space of all possible allele genealogies and allele frequency trajectories for the selected allele.

Although these likelihood-based methods achieve state-of-the-art results, a potential downside is that they are computationally expensive. The method of Stern, Wilton, and Nielsen, for example, depends on obtaining a posterior sample of local trees from the program ARGweaver (Rasmussen et al., 2014), which can take many hours to generate even for moderate sample sizes. In practice, this makes it less likely that such methods would be employed in the exploratory phase of an analysis, as is routinely done with e.g. Tajima's D. It seems that there is scope for a method that is easy to deploy while also mitigating some of the confounding issues described above.

2 Methods

Our starting point is the standard n-coalescent (Kingman, 1982a; Kingman, 1982b) which is defined as a stochastic process on the set of partitions of the set $\{1, \dots, n\}$. The process begins at time $t = 0$ in state $\mathcal{C}(0) = \{\{1\}, \dots, \{n\}\}$. The instantaneous transition rate at time t is $\binom{|\mathcal{C}(t)|}{2}$, where $1 \leq |\mathcal{C}(t)| \leq n$ denotes the number of blocks in the partition remaining at time t. When a transition occurs, the new state is obtained by choosing two partition blocks uniformly at random and merging them. Thus, the number of partition blocks decreases monotonically over time, continuing until it reaches the absorbing state $\{\{1, \dots, n\}\}$. The trajectory of states $\{\mathcal{C}(t) : t \geq 0\}$ can be straightforwardly identified with a bifurcating tree on n leaves, with internal nodes occuring upon each block merger. For this reason, Kingman's coalescent is often described as a distribution on binary trees.

An algorithm for drawing from Kingman's coalescent follows directly from the above description. It is listed in the supplement (Algorithm S1) for completeness, though it is quite well-known. In this paper, we focus on an equivalent, but less common, method of sampling from Kingman's coalescent, with the goal of obtaining a generalization which will prove useful for studying natural selection. This algorithm is shown in Algorithm 1. The

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

main distinction is that it proceeds *forwards* in time (i.e., from past up to present), as opposed to Kingman's original, retrospective process. That both the forwards- and backwards-in-time algorithms have the same distribution follows from e.g. Durrett (2008, Theorem 1.8).

2.1 The β-splitting family

We are motivated to consider Algorithm 1 because it can be generalized to produce alternative distributions on tree topologies. Observe that in line 5 of Algorithm 1, we could replace the uniform distribution with some other distribution on $\{1, \dots, |B_i|-1\}$. For example, a distribution which, for each $|B_i|$, placed mass $1/2$ on 1 and $|B_i|-1$, would produce unbalanced "caterpillar" trees with a large portion of external branches. Similarly, a distribution which placed all mass on (or near) $|B_i|/2$ would produce trees which tend to be more "balanced" than is observed under Kingman's coalescent. These two extremes produce the types of trees that we expect to form under certain types of natural selection, in particular directional and balancing selection.[2]

Such a model has been proposed by Aldous (1996), who studied probability distributions on random cladograms (topological trees with no branch length information). Aldous defined a one-parameter family of distributions which he called the *β-splitting model*[3]. In this model, a clade of size n is randomly split into subclades of sizes $\{i, n-i\}$, where now i is distributed according to a symmetric beta-binomial distribution with shape parameter β, conditioned on $i \notin \{0, n\}$. Concretely, this distribution is given by

$$p_n^\beta(i) = a_n^{-1}(\beta)\binom{n}{i}\int_0^1 x^i(1-x)^{n-i} f_\beta(x)\,\mathrm{d}x, \quad 1 \le i \le n-1, \tag{2}$$

where

$$f_\beta(x) \propto x^\beta (1-x)^\beta \tag{3}$$

is the symmetric beta density with shape parameter β, and

$$a_n(\beta) = \int_0^1 [1 - x^n - (1-x)^n] f_\beta(x)\,\mathrm{d}x$$

[2] A third type of selection, background selection, alters genetic diversity in a way that is indistinguishable from shrinking the effective population size (Charlesworth, Morgan, and Charlesworth, 1993), and is therefore not captured by our approach.

[3] This model should not be confused with the β-coalescent (Schweinsberg, 2003), which is a more general type of coalescent model that allows for multiple merger events. We discuss possible connections between generalized coalescent processes and our model in Section 4.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is the normalizing constant. Integrating out x in (2), one obtains

$$p_n^\beta(i) = a_n^{-1}(\beta)\frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{i!(n-i)!}, \quad 1 \le i \le n-1. \tag{4}$$

The beta density (3) is integrable for $\beta > -1$ (note that Aldous' parameterization differs by 1 from the usual convention.) However, up to normalization, (4) defines a valid probability distribution whenever $\min(\beta + i+1, \beta+n-1+1) > 0$; that is, for $\beta > -2$. For $\beta = 0$, $p_n(i;\beta) \propto 1$ and the distribution reduces to Kingman's coalescent. By examining the ratio

$$\frac{p_n^\beta(i)}{p_n^\beta(i+1)} = \frac{(i+1)(n+\beta-i)}{(\beta+i+1)(n-i)}, \tag{5}$$

we see that letting $\beta \to \infty$ causes p_n^β to place most of its mass near $n/2$, leading trees which are more "balanced" than under the usual coalescent. If $\beta \to -2$, ratio in (5) diverges for $i \in \{1, n-1\}$, so p_n^β places mass on $i \in \{1, n-1\}$, resulting in maximally unbalanced splits and a "caterpillar" tree.

The reader may wonder why the beta-binomial distribution was chosen, when we could conceivably have used any distribution on $\{1, \dots, n-1\}$. The symmetric beta-binomial is attractive due to parsimony (it adds only one extra parameter), and because it preserves some desirable properties of tree distributions such as exchangeability. Also, its usage has precedent in the related field of phylogenetics, where it has been proposed as a model for speciation (Blum and François, 2006). Other authors have recently studied further generalizations of this process to the case where the shape parameters are not symmetric (Sainudiin and Véber, 2016). A disadvantage of this model is that, in contrast to Kingman's coalescent, the forward-splitting process does not seem have any evolutionary interpretation (Aldous, 1996). We choose to view it empirically as a useful tool for studying natural selection using coalescent-based methods.

2.2 Expected site frequency spectrum

Given a sample of n individuals, the expected site frequency spectrum (ESFS) is the distribution of the number of individuals $i \in \{1, 2, \dots, n-1\}$ bearing the derived allele at a randomly selected segregating site. (We assume that the identity of the ancestral allele is known.) In this section we show how to deteremine the ESFS under the β-splitting model.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We denote the ESFS by $\mathbb{E}_\eta \boldsymbol{\xi}$, where the site frequency spectrum $\boldsymbol{\xi} \in \Delta^{n-1}$ is the sample version of ESFS, i.e. a vector whose i^{th} entry denotes the proportion of segregating sites where i members of the sample bear the derived allele. Here Δ^{n-1} denotes the $(n-1)$-dimensional probability simplex, i.e. the set of all numbers $x_1, \dots, x_n \geq 0$ such that $x_1 + \cdots + x_n = 1$. The expectation is taken with respect to genealogies generated under a given evolutionary model η. Although η could in principle be quite general, efficient methods for computing $\mathbb{E}_\eta \boldsymbol{\xi}$ are only known when η describes neutral evolution under either constant or variable effective population size. Therefore, from this point on we take η to represent a function representing the historical size of the population.

Under an "infinite sites" model with low rates of mutation, Griffiths and Tavaré (1998) have shown the following key result:

$$(\mathbb{E}_\eta \boldsymbol{\xi})_b \propto \sum_{k=2}^{n} p_{nkb} \cdot \mathbb{E}_\eta T_{nk}. \tag{6}$$

In the preceding display, $\mathbb{E}_\eta T_{nk}$ is the average amount of time (under the evolutionary model η) during which there are k lineages ancestral to a sample of size n, and p_{nkb} the probability that a branch a level k in an n-coalescent tree has b sampled descendants in the present.

In Kingman's coalescent,

$$p_{nkb} = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}, \tag{7}$$

which can be derived by a combinatorial "stars-and-bars" argument (Durrett, 2008). If the effective population size is constant, then $\mathbb{E}T_{nk} = \binom{k}{2}^{-1}$, from which follows the well known result that $(\mathbb{E}\boldsymbol{\xi})_b \propto 1/b$ for Kingman's coalescent. If population size varies through time according to some size history function $\eta(t)$, then a simple expression for $\mathbb{E}_\eta T_{nk}$ no longer exists, but Polanski and Kimmel (2003) have shown that it may be computed as a certain linear transformation of the vector of first coalescent times $\mathbb{E}_\eta T_{jj}$, $j = 2, \dots, n$. We return to this fact below.

Although Kingman's coalescent and its generalization to variable effective population size are the two best-known applications of Griffiths and Tavare's formula (6), in fact their argument holds more generally for any distribution on trees, assuming (crucially) that the branch lengths and topology of those trees are independent. Since this is true for the β-splitting model defined above, we can use a generalization of (6) to derive its expected SFS.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Let $\mathbb{E}_{(\beta,\eta)}$ denote expectation with respect to trees generated under the β-splitting model. Since the β-splitting model alters tree topology only, we have

$$(\mathbb{E}_{(\beta,\eta)}\boldsymbol{\xi})_b \propto \sum_{k=2}^{n} p^{\beta}_{nkb} \cdot \mathbb{E}_{\eta} T_{nk}, \tag{8}$$

where the vector $\mathbf{p}^{\beta}_{nk} = (p^{\beta}_{n,k,1}, \dots, p^{\beta}_{n,k,n-1})$ has the same interpretation as above. In the next two subsections, we show how to compute the "topological" (p^{β}_{nkb}) and "branch length" ($\mathbb{E}_{\eta} T_{nk}$) components of this formula.

2.2.1 Dynamic programming algorithm for $\mathbf{p}^{\beta}_{nk}$

A simple expression like (7) does not seem to exist when $\beta \neq 0$. Instead, we derive a dynamic programming algorithm for calculating the combinatorial factors $\mathbf{p}^{\beta}_{nk} \in \mathbb{R}^{n-1}, k = 2, \dots, n$ defined in the preceding section. The method applies to any forward-splitting model and includes β-splitting as a special case.

Define $f^{\beta}_{k,i,j}$ to be the probability that a size-i block at level k splits into blocks of size j and $i - j$. From the preceding section, we know that under Kingman's coalescent,

$$f^{(\beta=0)}_{k,i,j} \propto \frac{i-1}{n-k},$$

and for the general β-splitting model,

$$f^{\beta}_{k,i,j} \propto \frac{i-1}{n-k}\left[p^{\beta}_{i}(j) + p^{\beta}_{i}(i-j)\right]$$

where $p^{\beta}_{i}(\cdot)$ was defined in equation (4).

For each level k let $\mathbf{S}^k \in \mathbb{Z}^n$ be a row vector such that S^k_b is number of nodes at level k which subtend $b = 1, \dots, n$ leaves at the bottom of the coalescent tree. Also let $\mathbf{e}_1, \dots, \mathbf{e}_n \in \mathbb{R}^n$ be the standard basis (row-)vectors. Under the forward-splitting model described above, the sequence $\mathbf{S}^1, \mathbf{S}^2, \dots, \mathbf{S}^n$ forms a Markov chain, with transition probabilities

$$\mathbb{P}_{\beta}(\mathbf{S}^k = \mathbf{s} - \mathbf{e}_i + \mathbf{e}_j + \mathbf{e}_{i-j} \mid \mathbf{S}^{k-1} = \mathbf{s}) = \frac{i-1}{n-k+1} \cdot [f^{\beta}_{k-1,i,j} + f^{\beta}_{k-1,i,(i-j)}]. \tag{9}$$

The starting state of the Markov chain is $\mathbf{S}^1 = (0, 0, \dots, 1) = \mathbf{e}_n$. Focusing on an individual entry S^{k+1}_j and summing over all possible events that would cause it to *increase*, we obtain

$$\mathbb{P}(S^k_j = s_j + 1 \mid \mathbf{S}^{k-1} = \mathbf{s}) = \frac{1}{n-k+1} \sum_{\ell=j+1}^{n} (\ell - 1) s_\ell \sum_{q \in \{j, \ell-j\}} f^{\beta}_{k-1,\ell,q}. \tag{10}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Similarly, a *decrease* can occur only if a size-j block was chosen to split in the preceding level:

$$\mathbb{P}(S_j^k = s_j - 1 \mid \mathbf{S}^{k-1} = \mathbf{s}) = \frac{j-1}{n-k+1} s_j. \tag{11}$$

In matrix notation, (10) and (11) combine to yield

$$\mathbb{E}(\mathbf{S}^k \mid \mathbf{S}^{k-1} = \mathbf{s}) = \mathbf{s}\left(I_n + \frac{Q_{nk}^{\beta}}{n-k+1}\right)$$

where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix, and

$$\begin{aligned} Q_{nk}^{\beta} &= (F_{nk}^{\beta} - I_n)L_n \\ L_n &= \operatorname{diag}(0, 1, \dots, n-1) \\ F_{nk}^{\beta} &\in \mathbb{R}^{n\times n} \\ (F_{nk}^{\beta})_{i,j} &= f_{k,i,j}^{\beta} + f_{k,i,(i-j)}^{\beta}. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}(\mathbf{S}^k) &= \mathbb{E}(\mathbf{S}^{k-1})\left(I_n + \frac{Q_{nk}^{\beta}}{n-k+1}\right) \\ &= \cdots = \mathbb{E}(\mathbf{S}^1)\prod_{i=2}^{k}\left(I_n + \frac{Q_{ni}^{\beta}}{n-i+1}\right) \\ &= \mathbf{e}_n\prod_{i=2}^{k}\left(I_n + \frac{Q_{ni}^{\beta}}{n-i+1}\right). \end{aligned}$$

Finally,

$$\mathbf{p}_{nk}^{\beta} = \frac{1}{k}\mathbb{E}(\mathbf{S}^k).$$

2.2.2 Computing the expected branch lengths

Next we discuss how to compute the other necessary quantity $\mathbb{E}_\eta T_{nk}$ in equation (8). Let $\mathbf{T}_n = (T_{n,2}, T_{n,3}, \dots, T_{n,n})$ be the vector of these times. Polanski, Bobrowski, and Kimmel (2003) have shown the following relationship for a general size history function η:

$$\mathbb{E}_\eta \mathbf{T}_n = \mathbf{A} \cdot \mathbb{E}_\eta \tilde{\mathbf{T}}_n \tag{12}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $\mathbb{E}_\eta \tilde{\mathbf{T}}_n$ is the vector of first coalescent times,

$$\mathbb{E}_\eta \tilde{T}_{nj} = \int_0^\infty \exp\left\{-\binom{j}{2} R_\eta(t)\right\} \mathrm{d}t, \quad j = 2, \dots, n \tag{13}$$

$$R_\eta(t) := \int_0^t \frac{\mathrm{d}s}{\eta(s)},$$

and $\mathbf{A}_n \in \mathbb{R}^{(n-1)\times(n-1)}$ has entries

$$A_{n,k,j} = \frac{\prod_{l=k,l\neq j}^n \binom{l}{2}}{\prod_{l=k,l\neq j}^n \left[\binom{l}{2} - \binom{j}{2}\right]}.$$

As in the preceding section, this result holds for any tree distribution in which branch lengths and topology are independent, so it can be applied to our model.

Readers who are familiar with this area may notice that, for Kingman's coalescent, the expected SFS is typically not calculated via equation (8). Instead, by another result of Polanski and Kimmel (2003), interchanging the order of summations in equations (8) and (12) allows the (unnormalized) ESFS to be expressed as a linear transformation of $\mathbb{E}_\eta \tilde{\mathbf{T}}_n$. Unfortunately, this trick does not lead to simplifications in our more general model, so we first compute the expected intercoalescence times and then plug them into (8). For large n, the matrix-vector product (13) is numerically unstable, so we use a high precision numerical library to evaluate the integral (13) and then (12). This approach is less efficient than using hardware floating point operations, but it only needs to be performed once per given demography, so it is suitable for genomewide analysis.

2.3 Estimating β

Given the probabilistic model defined above, how can we estimate it in order to infer β? In this section, we propose two methods depending on the type of data that are available.

2.3.1 From the SFS

To perform inference using the SFS we rely on the so-called *Poisson random field* (PRF) approximation (Sawyer and Hartl, 1992), which assumes the coalescent tree at every segregating site is independent of all others.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Assuming also that mutations are rare—formally, that $\theta \to 0$, as is reasonable for humans and many other species—then we may approximate the mutation process on a coalescent tree by a Poisson process.

Given an empirical frequency spectrum $\boldsymbol{\phi} \in \mathbb{Z}^{n-1}$, where ϕ_i is the number of segregating sites where i copies of the derived allele were observed, the PRF log-likelihood is

$$L(\beta, \theta | \boldsymbol{\phi}) = ||\boldsymbol{\phi}||_1 \log(\theta \|\mathbb{E}_{\eta,\beta}\boldsymbol{\xi}\|_1) - \theta \|\mathbb{E}_{\eta,\beta}\boldsymbol{\xi}\|_1 + \frac{\langle \boldsymbol{\phi}, \mathbb{E}_{\eta,\beta}\boldsymbol{\xi}\rangle}{\|\mathbb{E}_{\eta,\beta}\boldsymbol{\xi}\|_1}, \tag{14}$$

where the ESFS $\mathbb{E}_{\eta,\beta}\boldsymbol{\xi}$ is calculated using the procedure derived in Section 2.2. If the mutation rate θ is not known, then the maximum likelihood estimate can be shown to equal

$$\hat{\theta}_{\mathrm{MLE}} = \frac{\|\boldsymbol{\phi}\|_1}{\|\mathbb{E}_{\eta,\beta}\boldsymbol{\xi}\|_1}.$$

Substituting this back into (14), and setting $\mathbf{p} = \boldsymbol{\phi}/\|\boldsymbol{\phi}\|_1$, $\mathbf{q}(\beta) = \mathbb{E}_{\eta,\beta}\boldsymbol{\xi}/\|\mathbb{E}_{\eta,\beta}\boldsymbol{\xi}\|_1$, we obtain that the profile likelihood

$$L(\beta|\boldsymbol{\phi}) = L(\beta, \hat{\theta}_{\mathrm{MLE}} \mid \boldsymbol{\phi}) = -D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}(\beta)) + \text{const.}$$

In order words, maximizing the likelihood is equivalent to minimizing the KL divergence between the categorical distributions $\mathbf{p}$ and $\mathbf{q}(\beta)$ (Bhaskar, Wang, and Song, 2015).

2.3.2 From inferred trees

The ESFS is obtained by integrating over all possible genealogies at a given site, and then fit to data by assuming independence between sites. An alternative strategy is try to estimate those genealogies, and then do inference conditioned on them. Recently in population genetics, there have been methodological breakthroughs that enable the estimation of ancestral recombination graphs using large numbers of genomes (Kelleher et al., 2019a; Speidel et al., 2019). In the future, as algorithms and computational capabilities continue to improve, this may become the dominant mode of population genetic analysis. We therefore explored extensions of our methods to the case where genealogies are estimated instead of integrated out.

Because of the probabilistic nature of our model, it is easy to extend it to the case where the genealogy is observed instead of latent. Moreover, estimating β conditional on a collection of inferred genealogies simplifies the problem considerably. If we assume a bifurcating tree, the sizes of children

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

nodes can be modeled by the beta-binomial distribution as previously described. Just like the preceding section, we proceed level by level in the (now observed) genealogy. At each level $k = 2, \ldots, n$ of the tree, let the size of the internal node which splits into two child nodes be denoted B_k, and the sizes of its child nodes c_k and $B_k - c_k$. We model the probability of an the observed tree $\mathcal{T}$ as

$$\mathbb{P}(\mathcal{T} \mid \beta) = \prod_{k=2}^{n} p_{B_k}^{\beta}(c_k), \tag{15}$$

with $p_{B_k}^{\beta}$ defined as in (4), so that $\hat{\beta}$ obtained by numerical optimization.

Weighted likelihood When experimenting with this method, we observed a small but consistent performance improvement by reweighting the likelihood (15):

$$\mathbb{P}(\mathcal{T} \mid \beta) = \prod_{k=2}^{n} [p_{B_k}^{\beta}(c_k)]^{w(k)},$$

where $w(k)$ is a weighting function. For detecting directional selection, we found that setting the weights proportional to the size of the internal node, $w(k) = B_k$, worked well. For detecting balancing selection, we found that it helped to weight the various terms by total amount of branch length at their respective level in the tree: $w(k) = kt_k$, where t_k is the amount of branch length at level k in the tree (see Section 3.1.) Using weights improved the method's performance of detecting the imbalance of the tree. The effect of the different weighting methods is shown in Figures S4 and S5. The gain was around 0.01–0.04 AUC in each scenario.

Related tree imbalance statistic The Colless statistic (Mooers and Heard, 1997) is a measure of the imbalance of a binary tree, defined as

$$I_c(\mathcal{T}) = \frac{1}{\binom{n-1}{2}} \sum_{t \in \mathcal{T}} |t_r - t_\ell|, \tag{16}$$

where the summation is over all internal nodes t of the tree, and t_r, t_ℓ are the sizes of the two child nodes descending from t. We used the Colless statistic as a baseline for comparing the performance of our $\hat{\beta}$ statistic when fitted to inferred trees. The exact relationship between $\hat{\beta}$ and $I_c(\mathcal{T})$ is somewhat opaque, but in general we can note that I_c is maximized for a caterpillar tree, and is zero for a perfectly balanced tree with an even number of leaves.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hence it is negatively associated with β-splitting parameter. In the next section, we compare the ability of these two measures to detect signals of selection.

Polytomies In practice, we found that current tree inference softwares often generate multifurcating trees. Since our method assumes a bifurcating tree, we first resolved these polytomies by arbitrarily breaking them into sequences of bifurcation events. Of course, polytomies could well represent additional selection signal. Our current implementation ignores this, but we discuss potential extensions in Section 4.

2.4 Alternative parameterization

We conclude this section with a note on implementation. When fitting our model to data, we observed that the parameterization (4) exhibited some numerical instability when performing gradient-based optimization. The problem arises when computing the normalizing constant for the range $-2 < \beta < -1$ which, as mentioned in Section 2.1, can no longer be interpreted as a draw from a conditioned beta-binomial distribution. To work around this, we restrict $\beta > -1$ and then perform a log transformation. Specifically, in all of the results reported below, the following alternative definition of the symmetric beta-binomial distribution is used:

$$\mathrm{BB}(i|n,\beta) = \frac{\Gamma(n+1)}{\Gamma(i+1)\Gamma(n-i+1)} \frac{\Gamma(i+e^{\beta})\Gamma(n-i+e^{\beta})}{\Gamma(n+2e^{\beta})} \frac{\Gamma(2e^{\beta})}{\Gamma(e^{2\beta})}$$

Then we restricted i to be in $\{1, 2, \ldots, n-1\}$;

$$p_n^{\beta}(i) = \frac{\mathrm{BB}(i|n,\beta)}{1 - \mathrm{BB}(0|n,\beta) - \mathrm{BB}(n|n,\beta)} \tag{17}$$

where $i \in \{0, 1, \ldots, n-1\}$, $n \in \mathbb{N}^{+}$ and $\beta \in \mathbb{R}$. The transformed distribution has the following properties: when $\beta = 0$, this becomes a uniform distribution so the model recovers the usual Kingman's Coalescent. When $\beta \to -\infty$, most of the weights of the distribution will be at the tails, so corresponding tree will be similar to a caterpillar tree. And when $\beta \to \infty$, the weights will be accumulated around the center and lead to a balanced tree.

2.5 Data analysis pipeline

A description of the pipeline used to analyze data and run our methods is contained in the supplement (Section S1).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3 Results

In this section, we study various characteristics of the methods we derived in Section 2 using simulations, before concluding with applications to real data.

3.1 Topological variance analysis

Recently Ferretti et al. (2017) gave an interpretation of several frequency spectrum-based neutrality tests in terms of tree imbalance. In this section we study our model using some of their results. This helps clarify the connection between some existing neutrality tests and our work.

Following Ferretti et al., we define d_k to be the size (number of leaf nodes subtended by) a randomly selected lineage at level k in a genealogy. Averaged over genealogies under the β-splitting model, we have

$$\begin{aligned}\operatorname{var}_\beta(d_k) &= \sum_{b=1}^{n-k+1} b^2 p^\beta_{nkb} - (\mathbb{E}_\beta d_k)^2 \\ &= \sum_{b=1}^{n-k+1} b^2 p^\beta_{nkb} - \left(\frac{n}{k}\right)^2,\end{aligned}$$

where $\mathbf{p}^\beta_{nk}$ was defined in Section 2.2.1, and the second inequality holds because $\mathbb{E}d_k = n/k$ under any leaf-exchangeable tree distribution. Computing $\operatorname{var}_\beta(d_k)$ in closed form for our model is challenging due to the fact that $\mathbf{p}^\beta_{nk}$ is recursively defined. Here we focus on a few special cases where we can derive a precise answer, and study the general relationship using simulations.

For $\beta \to -2$, corresponding to the caterpillar tree, it is easy to show that

$$\lim_{\beta\to-2} \operatorname{var}_\beta(d_k) = (k-1)\left(\frac{n}{k}-1\right)^2, \tag{18}$$

as already noted by Ferreti et al. Also, for Kingman's coalescent, $\beta = 0$,

$$\operatorname{var}_{\beta=0}(d_k) = \sum_{b=1}^{n-k+1} b^2 \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} - \left(\frac{n}{k}\right)^2 = \frac{n(n-k)(k-1)}{k^2(k+1)}. \tag{19}$$

For $\beta \to \infty$, we were unable to derive a closed-form expression for $\lim_{\beta\to\infty} \operatorname{var}_\beta(d_k)$. However, Ferretti et al. showed that the dominant contribution to topological variance comes from level $k = 2$, for which

$$\operatorname{var}_\beta(d_2) = \operatorname{var}(X \mid 1 \le X \le n-1), \text{ where } X \sim \operatorname{BetaBinomial}(n; \beta, \beta).$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

If n and β are both large, the condition $1 \leq X \leq n-1$ has probability near one and can be ignored. Using the variance formula for the beta-binomial distribution, we have

$$\lim_{\beta\to\infty} \mathrm{var}_\beta(d_2) \leq \lim_{\beta\to\infty} \frac{n\beta^2(n+2\beta)}{4\beta^2(2\beta+1)} \approx \frac{n}{4}, \text{ for large } n.$$

We further define

$$\overline{\mathrm{var}_\beta(d \mid \mathcal{T})} = \frac{1}{l}\sum_{k=2}^{n} kt_k \,\mathrm{var}_\beta(d_k),$$

which is the topological variance of a given genealogy, weighted by the relative proportion of branch length at each level (see equation (4) in Ferretti et al.). Substituting t_k and l by their expected values in equations (18) and (19), as $n \to \infty$,

$$\begin{aligned} \overline{\mathrm{var}_{\beta=0}(d)} &= H_{n-1}^{-1}\sum_{k=2}^{n}\frac{n(n-k)}{k^2(k+1)} \asymp \frac{\pi^2-9}{6}\cdot\frac{n^2}{\log n} \\ \lim_{\beta\to-2}\overline{\mathrm{var}_\beta(d)} &= H_{n-1}^{-1}\sum_{k=2}^{n}\left(\frac{n}{k}-1\right)^2 \asymp \frac{\pi^2-6}{6}\cdot\frac{n^2}{\log n}, \end{aligned}$$

where H_n is the nth harmonic number.

Now let T be a neutrality test statistic (for example, Tajima's D or Fay and Wu's H). Since the parameter β only affects tree topology, we obtain from formula (17) of Ferretti et al.,

$$\mathbb{E}_\beta T = \mathbb{E}_\beta T - \mathbb{E}_{\beta=0} T = \alpha_T^n\left(\overline{\mathrm{var}_\beta(d)} - \overline{\mathrm{var}_{\beta=0}(d)}\right),$$

where $\alpha_T(n)$ is a test-specific constant which depends on n, and for simplicity we ignored the normalization term $f_{\mathbf{\Omega}}(\theta l)$.

To show an example of how the topological variance affects neutrality tests such as Tajima's D, we simulated genealogies under various settings of β, assuming constant population size with no recombination (Figure 1). The box plots are empirical distributions of two neutrality tests (Tajima's D and Fay and Wu's H) for various settings of $\beta \in [-2, \infty)$. The dashed red lines represent the limiting values predicted by the calculations shown above. The figure shows how to different values of these statistics can be interpreted in terms of β, and vice versa. We see, for example, that D and H appear to be more sensitive to $\beta < 0$, in the sense that their distribution at $\beta = 0$ nearer the $\beta \to \infty$ limit than the $\beta \to -2$ limit.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2 Simulated data

To benchmark our methods on simulated data, we studied their ability to classify simulated genomic regions as being either neutral or under some form of selection. The receiver operating characteristic (ROC) curve, and associated area under curve (AUC) statistic, are standard ways to measure the performance of a classifier. For each experiment described below, we generated data under two different models, and then plotted ROC curves for each method. The two possible models are printed at the top of each ROC curve. The legend lists each method that was compared, along with its AUC score.

The classification procedures derived from our methods are denoted btree and bsfs. The bsfs results were obtained by maximizing (14) over β with respect to the observed frequency spectrum. btree is the tree-sequence based estimate, obtained by maximizing the conditional likelihood defined in (15) over β conditional on a given tree. As a baseline, we also compared our method to Colless' statistic (see Section 2.3.2) and Tajima's D. Finally, ROC curves were computed by thresholding the empirical null distributions of each test statistic. We also use these neutrally evolved simulations to infer population size histories ($\eta(t)$) that we use for bsfs. Our simulation process is explained in detail in Section S1.1.

3.2.1 Directional selection

We simulated a single population with constant population size $N = 2 \times 10^4$. The simulated region was 10^5 base pairs, with recombination and mutation rates of 1.25×10^{-8} and 2.5×10^{-8} per base pair per generation, respectively. When each simulation terminated, we randomly sampled $n = 50$ haploid genomes and computed the relevant test statistics. We introduced a beneficial mutation 250 generations prior to present into the middle of the region, and we restarted the simulation if the mutation is lost or fixed. Following Stern, Wilton, and Nielsen (2019), we varied two parameters; selection coefficient $s \in \{.001, .003, .01, .02\}$ and allele frequency $F \in \{0.25, 0.5, 0.75\}$ of mutation when the simulation terminated. Genic selection was assumed, i.e. the relative fitnesses of the wild-type homozygotes, heterzygotes, and derived homozygotes were 1, $1 + s/2$, and $1 + s$, respectively.

Figure 2 displays results for each of the methods. In general, we observed that tree-sequence based methods are better at detecting strong selection compared to SFS-based methods. This is expected, because a recent hard sweep leaves a signal of elevated linkage disequilibrium that is invisible in

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the frequency spectrum (Kaplan, Hudson, and Langley, 1989). In particular, the btree method achieves at least 0.8 AUC for $s \geq 0.003$ and $F \geq 0.5$. The performance Colless' statistic and btree are similar. btree has significantly higher AUC ($p = .014$, Wilcoxon signed rank test), but the overall gain is small (mean $\Delta\text{AUC} = 0.0034$). Among the SFS-based statistics, our method (bsfs) achieved significantly higher AUC scores ($p = .0014$, Wilcoxon signed rank test) than Tajima's D, and the average gain is notable (mean $\Delta\text{AUC} = 0.049$).

3.2.2 Balancing selection

Next we studied our methods' ability to detect long-term balancing selection. Since this type of selection acts on a longer time scale than directional selection (Charlesworth, 2006), it is necessary to forward simulate for many more generations. To speed up the simulations, we reduced the population size by a factor of 10 to $N = 2 \times 10^3$, and increased the mutation and recombination rates to 1.25×10^{-7} and 2.5×10^{-7}. The simulated region was 2500 base pairs. When each simulation terminated we randomly sampled $n = 250$ haploid genomes and computed the relevant test statistics using them. Heterozygously advantageous mutations were introduced at constant rate throughout the simulation. We varied two parameters: $t_0 \in \{2 \times 10^3, 3 \times 10^3, 4 \times 10^3, 5 \times 10^3\}$ which represents the number of generations before present when beneficial mutations began, and selection coefficient $s \in \{.0004, .0008, .002\}$. The dominance parameter was set to $h = 25$ in all cases. Thus the fitnesses of the homo- and heterozygote were ≈ 1 and $s \cdot h \in \{.01, .02, .05\}$, respectively.

Figure 3 contains ROC curves along with the AUC values in parenthesis for each of the methods. For balancing selection btree again outperforms Colless' statistic, but the difference is subtle (mean $\Delta\text{AUC} = 0.0026$) and not significant ($p = 0.31$, Wilcoxon signed rank test). In contrast to the case of directional selection, SFS-based statistics did better than tree-based statistics in this example. Among the SFS-based statistics, our method (bsfs) achieved significantly higher AUC scores ($p = 0.0011$, Wilcoxon signed rank test) than Tajima's D with a mean $\Delta\text{AUC} = 0.024$. We performed some additional analysis to better understand why SFS-based statistics are better than the tree-based ones for detecting balancing selection. We found that long branches near the root of the tree that occur in genealogies under long-term balancing selection have a pronounced impact on the SFS, but do not affect the topology of inferred trees.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2.3 Effect of variable population size

It is well known that, when used to detect natural selection, Tajima's D is confounded by population structure and changes in effective population size (Stajich and Hahn, 2005; Biswas and Akey, 2006). In the single-population case, one interpretation of this phenomenon is that D measures both topological and branch length distortions compared to the neutral coalescent (Ferretti et al., 2017), and population size changes also distort branch lengths. In contrast, our SFS-based estimator is designed to detect topological changes only, and it can be modified to take into account population size history (Section 2.2).

We compared the ability of D and bsfs to detect directional selection under four scenarios:

- Constant population size under neutrality;
- Exponential growth under neutrality;
- Constant population size with directional selection;
- Exponential growth directional selection.

For the selective scenarios, we introduced a single mutation 250 generations before present to the middle of the 10^5 base pair region, restarting the simulation if the mutation was lost or fixed. The sample size was $n = 250$ haploids. The recombination and mutation rates were again 1.25×10^{-7} and 2.5×10^{-7}. For the bsfs method, we first estimated the underlying population size history $\eta(t)$ using 25Mb of neutral data simulated under the corresponding demography. Other varying parameters for the experiments can be seen at Table S1. In the table, $N_e(0)$ is the population size at the time simulation starts, g is the growth rate of exponential growth, s is the selective coefficient of the beneficial mutation and h is the dominance parameter.

In Figure 4a, our method has higher AUC than D for distinguishing a neutral model from selection for both constant population size and exponential growth (left and center panels). To illustrate the pitfalls of using D without correcting for demography, we also considered a third scenario (right-most panel) in which there is *no* selection; the only difference between the two models is that one of them underwent exponential growth, while effective population size in the other was constant. In this plot, a "true positive" signifies that the constant-sized model is rejected in favor of the exponential growth model when the latter model generated the data,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and similarly for a false positive. As expected, the plot shows that D has high power to detect exponential growth—however, if the analyst were unaware that the population had experienced growth, then this could wrongly be interpreted as evidence for selection. In contrast, after adjusting the expected frequency spectrum to compensate for this effect, our estimator does no better than a coin-toss (AUC ≈ 0.5) at distinguishing between the two régimes.

Another way to see this result is in Figure 4b, which shows the empirical distributions of D and $\hat{\beta}$ obtained from bsfs. After correcting for demography, the two neutral simulations (orange and blue) have roughly the same empirical distribution using our method, even though they are generated under quite different growth models. In contrast, the distribution of D under neutral exponential growth closely matches that of directional selection under exponential growth, and is very different from the distribution under neutrality and constant population size.

We repeated this experiment under simulated balancing selection. We again simulated four different scenarios:

1. Constant size with no advantageous mutation;
2. Exponential growth with no advantageous mutation;
3. Constant size with heterozygote advantage and;
4. Exponential growth with heterozygote advantage.

For the exponential growth scenarios, the growth began 250 generations ago. Detailed settings for each type of simulation are shown in Table S2. Results were similar to the directional selection experiment. In Figure 5b, we see that selection and growth "cancel out" in Tajima's D: it has a similar distribution under exponential growth and balancing selection as under neutrality with constant size. In contrast, the null distribution of bsfs is invariant after correcting for demography.

3.3 Real data analysis

We applied our models to data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), using tree sequences that were inferred by Kelleher et al. (2019b). To understand how our model works compared to other known statistics, we focused on 7 regions which are known to experience selection: *LCT* in chromosome 2, *SLC45A2* in chromosome 5, *HERC2* in chromosome 15 for European populations; *SLC44A5* in chromosome 1,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

EDAR in chromosome 2, *ADH1* in chromosome 4 for East Asian populations; *MHC* in chromosome 6 for all populations. For *LCT*, *SLC45A2*, *HERC2*, *SLC44A5*, *EDAR*, *ADH1* we used the btree statistic to investigate directional selection since it is sensitive to linkage disequilibrium. For *MHC* we used bsfs since our simulation results show that our frequency spectrum-based methods are better at detecting long-term balancing selection. We performed one-sided testing: for directional selection, p-values were calculated by p^-, and for balancing selection by p^+ (cf. eqn. 21).

3.3.1 Directional selection

Lactose is the principle sugar in milk. Like other mammals, humans historically lost the intestinal enzyme lactase after infancy, and with it the ability to digest milk. But between 5,000 to 10,000 years ago, a genetic mutation arose that confers lactase persistence in adults. Today it is found in a majority of the adult populations of Northern and Central Europe. The location of this mutation in the gene *LCT* displays one of the strongest signals of directional selection in the human genome (Bersaglieri et al., 2004).

In Figure 6a, as expected we have a very small p-value for the European populations around *LCT*. This indicates our estimated β-splitting parameters are negative, as expected for strong directional selection (Section 2.1). Specifically, Utah Residents with Northern and Western European Ancestry (CEU), British in England and Scotland (GBR) and Finnish in Finland (FIN) have significantly negative $\hat{\beta}$. Southern European populations such as Toscani in Italia (TSI) and Iberian Population in Spain (IBS) also show evidence of selection, though the signal is weaker, reflecting the fact that the strength of selection may be lower in these populations (Gerbault et al., 2011).

SLC45A2 is a gene related to pigmentation (Branicki et al., 2008). It encodes a transporter protein that mediates melanin synthesis. In humans, it has been identified as a factor in the light skin of Europeans. As shown in Figure 6b, selection signals tended to be noisier in this region, and our median centered btree statistic does not see a pronounced peak this gene. The segments around this gene have small p-values for only TSI and CEU. However, the p-values are not above the genome-wide Bonferroni threshold, and are eclipsed by other nearby regions.

Figure 6c shows results for *HERC2*, which is associated with eye and skin pigmentation (Donnelly et al., 2012). Around this region there are blue-eye associated alleles found at high frequencies in European populations. In our results, the lowest p-value belongs to FIN, followed by GBR and CEU.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Turning to East Asian populations, we first studied *SLC44A5*, which is associated with neurological diseases and has been reported in several recent papers to be under selection in Japanese and Chinese populations (Liu et al., 2013; Zhao et al., 2019; Yasumizu et al., 2020). Our method confirms these findings (Figure 6), with highly significant hits centered on this gene for Japanese in Tokyo, Japan (JPT) and Han Chinese in Beijing, China (CHB).

We also found significant hits for all East Asian populations near *EDAR* (Figure 6e), again confirming earlier studies (Botchkarev and Fessing, 2005; Hlusko et al., 2018).

Finally, we examined the *ADH1* family. Alcohol is degraded primarily by alcohol dehydrogenase, and genetic variation affecting the rate of alcohol degradation found at *ADH1B* and *ADH1C*. Variants of these genes are thought to be associated with alcohol drinking habits and alcoholism. Our results (Figure 6f) confirm earlier findings (Han et al., 2007) that this family is under directional selection in Kinh in Ho Chi Minh City, Vietnam (KHV); Japanese in Tokyo, Japan (JPT); and Southern Han Chinese (CHS).

Estimates of the raw $\hat{\beta}$ values corresponding to these Manhattan plots are given in the supplement (see Figures S6 and S7).

3.3.2 Balancing selection

Next, we used our method to study long-term balancing selection in the the major histocompatibility complex (MHC). MHC is a large region of the vertebrate genome with immune-related functionality. Because evolution favors allelic diversity in this region (Takahata, 1993), we expect to detect signals of balancing selection in all populations. Our results (Figure 7) confirm this expectation; we observed highly significant signals across all 1000 Genomes subpopulations. Importantly, since this is an upper tail test for bsfs, we reject the null hypothesis that $\beta = 0$ in favor of the alternative $\beta > 0$. Thus, our method correctly infers that *MHC* is under balancing selection.

3.3.3 Results of genome-wide scan

In Section S5, we list the genomewide top hits (in terms of p-value) for the five major superpopulations in the 1000 Genomes dataset. They include a number of loci that are known to be under selection; such as *LCT*, *ALDH*; the *HLA* complex; and various pigmentation, and eye color-related genes. There are also other hits that, as far as we can tell, have not yet been

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

implicated by natural selection. Note that, due to linkage, many more genes are tagged than are likely under selection, but the genes should be proximal to a selected locus. A browser which can be used to explore all of our results, and compare them with classical tests of neutrality, is provided at the URL shown below.

4 Discussion

In this paper, we presented some new methods to detect natural selection using a generalization of Kingman's coalescent to the case where genealogies exhibit systematic topological imbalance. We showed how this leads to relatively simple estimators of selection that can be applied to frequency spectrum data, or just as easily to sequences of estimated genealogies. An important feature of our method is its ability to incorporate demographic information. Using simulations, we recapitulated the tendency, already well known in the literature, of widely used deviance statistics like Tajima's D to conflate variations in effective population size with natural selection. We showed that our method can correct for this tendency, by incorporating demographic estimates into its generative model of tree formation.

Our method is an example, albeit a basic one, of a recent trend towards likelihood-based methods for inferring natural selection from polymorphism data. We stress that our method will generally not be as sensitive as more elaborate and correct approximations to the coalescent under selection—compare, for example, the results of our Figures 2 and 3 with Figures 3 and 4 of Stern, Wilton, and Nielsen (2019). However, an advantage of our method is easy to understand and interpret, and also fast, requiring only to solve a univariate optimization problem. This can be done in only fractions of a second even for large sample sizes (Figure S3). Running our method on the entire 1000 Genomes dataset takes a few hours on a cluster. We see our work as adding to the toolbox of exploratory procedures that the analyst performs when studying a new dataset. Large "hits" yielded from our method can be used to flag a region for subsequent analysis, perhaps using more advanced and computationally expensive full-likelihood procedures. To this end, we have created an open source software package that makes it easy to run our methods. Researchers can also access our 1000 Genomes Project results with the browser we developed for this purpose. It enables to search through whole genome hits of our β estimates along with classical neutrality tests across for all populations.

There are several ways our model could be improved. We focused on

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the beta-binomial distribution because of its earlier usages in phylogenetics. However, as noted earlier, other distributions are possible, and perhaps some other model produces tree topology distributions that are more suited to studying natural selection. Another obvious criticism of our model is that it assumes that β is constant over time. This seems most appropriate for highly variable regions like *HLA*, where there is a continual introduction of new selected alleles. For regions that came under sudden directional selection as the result of the introduction of a beneficial allele, it would be better to use a model where the topological distribution of subtrees varies over time. This could allow for estimating the age of a selected variant, or understanding whether selection occurred on standing variation or because of the introduction of a new allele, both topics of longstanding interest in population genetics (Malaspinas et al., 2012; Hedrick, 2013; Barrett and Schluter, 2008; Feder, Kryazhimskiy, and Plotkin, 2014; Terhorst, Schlötterer, and Song, 2015; Palamara et al., 2018). Incorporating this feature into our SFS-based model would be challenging, as it creates dependence between the "time" and "topology" components of the expected frequency spectrum, thus invalidating equation (6). But it is easily added to the tree-based estimator in Section 2.3.2. We experimented with this, but found that the branch length estimates the current generation of tree sequence estimation programs are not yet reliable enough to support this kind of inference. As these methods continue to improve, this could be a future extension of our work.

When running our method on tree sequence data, we observed that estimated trees contained many polytomies. Since trees generated under Kingman's coalescent are almost surely bifurcating, we broke these polytomies arbitrarily in order to perform inference. However, the presence polytomies in estimated trees could be another signal of selection, particularly in the case of recent positive selection. Incorporating a probabilistic model of node size into our method could potentially make use of this signal. The Λ-coalescent (Sagitov, 1999; Pitman, 1999) is a generalization of Kingman's coalescent which allows for various forms of multiple-merger events. Research on inference methods under generalized coalescents is ongoing (Spence, Kamm, and Song, 2016; Blath et al., 2016). In the future, our method could be extended to work under this more general model.

Acknowledgements

This project was conceived during the workshop *Ecosystem dynamics: stakes, data and models* at Université Paris–Saclay. JT thanks the Institut Pascal

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for organizing the event, travel funding, and the limitless supply of M&M's; and Amandine Véber and Raazesh Sainudiin for helpful early discussions. JT was supported by NSF grant DMS-2052653.

Data and code availability

All of the data analyzed in this paper are publicly available. An open source implementation of our methods is available at https://github.com/jthlab/bim. Notebooks which reproduce our analyses are available at https://github.com/jthlab/bim-paper.

References

[1] Guillaume Achaz. "Frequency spectrum neutrality tests: one for all and all for one". In: *Genetics* 183.1 (2009), pp. 249–258.

[2] David Aldous. "Probability Distributions on Cladograms". en. In: *Random Discrete Structures*. Ed. by David Aldous and Robin Pemantle. The IMA Volumes in Mathematics and its Applications. New York, NY: Springer, 1996, pp. 1–18. ISBN: 978-1-4612-0719-1. DOI: 10.1007/978-1-4612-0719-1_1.

[3] Rowan DH Barrett and Dolph Schluter. "Adaptation from standing genetic variation". In: *Trends in ecology & evolution* 23.1 (2008), pp. 38–44.

[4] Jeremy J Berg and Graham Coop. "A coalescent model for a sweep of a unique standing variant". In: *Genetics* 201.2 (2015), pp. 707–725.

[5] Todd Bersaglieri et al. "Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene". en. In: *The American Journal of Human Genetics* 74.6 (June 2004), pp. 1111–1120. ISSN: 0002-9297. DOI: 10.1086/421051. URL: http://www.sciencedirect.com/science/article/pii/S0002929707628389 (visited on 12/26/2020).

[6] A. Bhaskar, Y. X. Rachel Wang, and Y. S. Song. "Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data". In: *Genome Research* 25.2 (2015), pp. 268–279.

[7] Shameek Biswas and Joshua M Akey. "Genomic insights into positive selection". In: *TRENDS in Genetics* 22.8 (2006), pp. 437–446.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[8] Jochen Blath et al. “The site-frequency spectrum associated with Ξ-coalescents”. In: *Theoretical Population Biology* 110 (2016), pp. 36–50.

[9] Michael GB Blum and Olivier François. “Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance”. In: *Systematic Biology* 55.4 (2006), pp. 685–691.

[10] Vladimir A. Botchkarev and Michael Y. Fessing. “Edar Signaling in the Control of Hair Follicle Development”. English. In: *Journal of Investigative Dermatology Symposium Proceedings* 10.3 (Dec. 2005). Publisher: Elsevier, pp. 247–251. ISSN: 1087-0024. DOI: 10.1111/j.1087-0024.2005.10129.x. URL: https://www.jidsponline.org/article/S0022-202X(15)52599-2/abstract (visited on 04/20/2021).

[11] Wojciech Branicki et al. “Association of the SLC45A2 gene with physiological human hair colour variation”. en. In: *Journal of Human Genetics* 53.11 (Dec. 2008). Number: 11 Publisher: Nature Publishing Group, pp. 966–971. ISSN: 1435-232X. DOI: 10.1007/s10038-008-0338-3. URL: https://www.nature.com/articles/jhg2008124 (visited on 04/20/2021).

[12] James J Cai et al. “Pervasive hitchhiking at coding and regulatory sites in humans”. In: *PLoS genetics* 5.1 (2009), e1000336.

[13] A. Celisse et al. “New efficient algorithms for multiple change-point detection with reproducing kernels”. en. In: *Computational Statistics & Data Analysis* 128 (Dec. 2018), pp. 200–220. ISSN: 0167-9473. DOI: 10.1016/j.csda.2018.07.002. URL: https://www.sciencedirect.com/science/article/pii/S0167947318301683 (visited on 04/19/2021).

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

[15] Deborah Charlesworth. “Balancing selection and its effects on sequences in nearby genome regions”. In: *PLoS Genet* 2.4 (2006), e64.

[16] Michael P. Donnelly et al. “A global view of the OCA2-HERC2 region and pigmentation”. In: *Human Genetics* 131.5 (2012), pp. 683–696. ISSN: 0340-6717. DOI: 10.1007/s00439-011-1110-x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3325407/ (visited on 04/20/2021).

[17] R. Durrett. *Probability Models for DNA Sequence Evolution*. 2nd. Springer, New York, 2008.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[18] J C Fay and C I Wu. "Hitchhiking under positive Darwinian selection". In: *Genetics* 155 (2000), pp. 1405–1413.

[19] Alison F Feder, Sergey Kryazhimskiy, and Joshua B Plotkin. "Identifying Signatures of Selection in Genetic Time Series". In: *Genetics* 196.2 (Feb. 2014), pp. 509–522.

[20] Luca Ferretti et al. "Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests". In: *Genetics* 207.1 (2017), pp. 229–240.

[21] Yair Field et al. "Detection of human adaptation during the past 2000 years". en. In: *Science* 354.6313 (Nov. 2016). Publisher: American Association for the Advancement of Science Section: Report, pp. 760–764. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aag0776. URL: https://science.sciencemag.org/content/354/6313/760 (visited on 04/27/2021).

[22] Yun-Xin Fu and Wen-Hsiung Li. "Statistical tests of neutrality of mutations." In: *Genetics* 133.3 (1993), pp. 693–709.

[23] Pascale Gerbault et al. "Evolution of lactase persistence: an example of human niche construction". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1566 (Mar. 2011), pp. 863–877. ISSN: 0962-8436. DOI: 10.1098/rstb.2010.0268. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3048992/ (visited on 04/28/2021).

[24] R.C. Griffiths and Simon Tavaré. "The age of a mutation in a general coalescent tree". In: *Communications in Statistics. Stochastic Models* 14.1-2 (1998), pp. 273–295.

[25] Benjamin C Haller and Philipp W Messer. "SLiM 3: Forward Genetic Simulations Beyond the WrightFisher Model". In: *Molecular Biology and Evolution* 36.3 (Mar. 2019), pp. 632–637. ISSN: 0737-4038. DOI: 10.1093/molbev/msy228. URL: https://doi.org/10.1093/molbev/msy228 (visited on 04/05/2021).

[26] Benjamin C Haller et al. "Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes". In: *Molecular ecology resources* 19.2 (2019), pp. 552–566.

[27] Yi Han et al. "Evidence of Positive Selection on a Class I ADH Locus". In: *American Journal of Human Genetics* 80.3 (Mar. 2007), pp. 441–456. ISSN: 0002-9297. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1821113/ (visited on 12/27/2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[28] Philip W Hedrick. "Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation". In: *Molecular ecology* 22.18 (2013), pp. 4606–4618.

[29] Leslea J. Hlusko et al. "Environmental selection during the last ice age on the mother-to-infant transmission of vitamin D and fatty acids through breast milk". en. In: *Proceedings of the National Academy of Sciences* 115.19 (May 2018). Publisher: National Academy of Sciences Section: PNAS Plus, E4426–E4432. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1711788115. URL: https://www.pnas.org/content/115/19/E4426 (visited on 04/20/2021).

[30] N. L. Kaplan, T. Darden, and R. R. Hudson. "The Coalescent Process in Models With Selection". In: *Genetics* 120 (1988), pp. 819–829.

[31] N. L. Kaplan, R. R. Hudson, and C. H. Langley. "The "hitchhiking effect" revisited". In: *Genetics* 123 (1989), pp. 887–899.

[32] Jerome Kelleher et al. "Inferring whole-genome histories in large population datasets". In: *Nature Genetics* 51.9 (2019), pp. 1330–1338.

[33] Jerome Kelleher et al. *Inferring whole-genome histories in large population datasets: inferred tree sequences for 1000 Genomes.* Version 1.0.0. May 2019. DOI: 10.5281/zenodo.3051855.

[34] Yuseob Kim and Rasmus Nielsen. "Linkage Disequilibrium as a Signature of Selective Sweeps". In: *Genetics* 167.3 (July 2004), pp. 1513–1524.

[35] J. F. C. Kingman. "On the genealogy of large populations". In: *J. Appl. Prob.* 19A (1982), pp. 27–43.

[36] J. F. C. Kingman. "The coalescent". In: *Stoch. Process. Appl.* 13 (1982), pp. 235–248.

[37] Stephen M. Krone and Claudia Neuhauser. "Ancestral processes with selection". In: *Theoretical Population Biology* 51.3 (1997), pp. 210–237.

[38] Erich L Lehmann and George Casella. *Theory of point estimation.* Springer Science & Business Media, 2006.

[39] Xuanyao Liu et al. "Detecting and Characterizing Genomic Signatures of Positive Selection in Global Populations". en. In: *The American Journal of Human Genetics* 92.6 (June 2013), pp. 866–881. ISSN: 00029297. DOI: 10.1016/j.ajhg.2013.04.021. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002929713001821 (visited on 04/20/2021).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[40] Kirk E Lohmueller et al. "Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome". In: *PLoS Genet* 7.10 (2011), e1002326.

[41] A. S. Malaspinas et al. "Estimating allele age and selection coefficient from time-serial data". In: *Genetics* 192.2 (2012), pp. 599–607.

[42] J. Maynard Smith and J. Haigh. "The hitch-hiking effect of a favourable gene". In: *Genet. Res., Camb.* 23 (1974), pp. 23–35.

[43] Graham McVicker et al. "Widespread genomic signatures of natural selection in hominid evolution". In: *PLoS Genet* 5.5 (2009), e1000471.

[44] A. Mooers and Stephen Heard. "Inferring Evolutionary Process from Phylogenetic Tree Shape". In: *Quarterly Review of Biology* 72 (Mar. 1997), pp. 31–54. DOI: 10.1086/419657.

[45] C Neuhauser and S M Krone. "The genealogy of samples in models with selection". In: *Genetics* 145 (1997), pp. 519–34.

[46] Jerzy Neyman and Egon Sharpe Pearson. "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[47] Pier Francesco Palamara et al. "High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability". In: *Nature Genetics* 50.9 (2018), pp. 1311–1317.

[48] Jim Pitman. "Coalescents with multiple collisions". In: *Annals of Probability* 27 (1999), pp. 1870–1902.

[49] Andrzej Polanski, Adam Bobrowski, and Marek Kimmel. "A note on distributions of times to coalescence, under time-dependent population size". In: *Theoretical Population Biology* 63.1 (2003), pp. 33–40.

[50] Andrzej Polanski and Marek Kimmel. "New Explicit Expressions for Relative Frequencies of Single-Nucleotide Polymorphisms With Application to Statistical Inference on Population Growth". In: *Genetics* 165.1 (2003), pp. 427–436.

[51] Matthew D Rasmussen et al. "Genome-wide inference of ancestral recombination graphs". In: *PLoS Genetics* 10.5 (2014), e1004342.

[52] Pardis C Sabeti et al. "Positive natural selection in the human lineage". In: *Science* 312.5780 (2006), pp. 1614–20.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[53] Serik Sagitov. "The general coalescent with asynchronous mergers of ancestral lines". In: *Journal of Applied Probability* 36.4 (1999), pp. 1116–1125.

[54] Raazesh Sainudiin and Amandine Véber. "A Beta-splitting model for evolutionary trees". In: *Royal Society open science* 3.5 (2016), p. 160016.

[55] Stanley A. Sawyer and Daniel L. Hartl. "Population genetics of polymorphism and divergence." In: *Genetics* 132.4 (1992), pp. 1161–1176.

[56] Jason Schweinsberg. "Coalescent processes obtained from supercritical Galton–Watson processes". In: *Stochastic processes and their Applications* 106.1 (2003), pp. 107–139.

[57] Leo Speidel et al. "A method for genome-wide genealogy estimation for thousands of samples". In: *Nature Genetics* 51.9 (2019), pp. 1321–1329.

[58] Jeffrey P Spence, John A Kamm, and Yun S Song. "The site frequency spectrum for general coalescents". In: *Genetics* 202.4 (2016), pp. 1549–1561.

[59] Jason E Stajich and Matthew W Hahn. "Disentangling the effects of demography and selection in human history". In: *Molecular Biology and Evolution* 22.1 (2005), pp. 63–73.

[60] Aaron J Stern and Rasmus Nielsen. "Detecting natural selection". In: *Handbook of Statistical Genomics: Two Volume Set* (2019), pp. 397–40.

[61] Aaron J. Stern, Peter R. Wilton, and Rasmus Nielsen. "An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data". en. In: *PLOS Genetics* 15.9 (Sept. 2019). Publisher: Public Library of Science, e1008384. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1008384. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008384 (visited on 04/05/2021).

[62] Fumio Tajima. "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." In: *Genetics* 123.3 (1989), pp. 585–595.

[63] N. Takahata. "Allelic genealogy and human evolution". In: *Mol. Biol. Evol.* 10 (1993), pp. 2–22.

[64] Jonathan Terhorst, Christian Schlötterer, and Yun S Song. "Multilocus analysis of genomic time series data from experimental evolution". In: *PLoS Genet* 11.4 (2015), e1005069.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[65] The 1000 Genomes Project Consortium. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74.

[66] Charles Truong, Laurent Oudre, and Nicolas Vayatis. "Selective review of offline change point detection methods". en. In: *Signal Processing* 167 (Feb. 2020), p. 107299. ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2019.107299. URL: https://www.sciencedirect.com/science/article/pii/S0165168419303494 (visited on 04/19/2021).

[67] Joseph J Vitti, Sharon R Grossman, and Pardis C Sabeti. "Detecting natural selection in genomic data". In: *Annual review of genetics* 47 (2013), pp. 97–120.

[68] B. F. Voight et al. "A Map of Recent Positive Selection in the Human Genome". In: *PLoS Biology* 4 (2006), e72.

[69] Yoshiaki Yasumizu et al. "Genome-Wide Natural Selection Signatures Are Linked to Genetic Risk of Modern Phenotypes in the Japanese Population". In: *Molecular Biology and Evolution* 37.5 (May 2020), pp. 1306–1316. ISSN: 0737-4038. DOI: 10.1093/molbev/msaa005. URL: https://doi.org/10.1093/molbev/msaa005 (visited on 04/20/2021).

[70] Bingxin Zhao et al. "Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits". en. In: *Nature Genetics* 51.11 (Nov. 2019). Number: 11 Publisher: Nature Publishing Group, pp. 1637–1644. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0516-6. URL: https://www.nature.com/articles/s41588-019-0516-6 (visited on 04/20/2021).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

```
Initialize $\mathcal{C}_1 = \{\{1, \ldots, n\}\}, k = 1$.
while $k < n$ do
    Sample $T_{k+1} \sim \text{Exp}(k(k+1)/2)$.
    Sample $B_i$ from $\mathcal{C}_k = \{B_1, \ldots, B_k\}$ with probability proportional
      to $(|B_i| - 1)/(n - k)$.
    Sample $m \sim \text{Uniform}(\{1, 2, \ldots, |B_i| - 1\})$.
    Randomly partition $B_i$ into non-empty subsets $A, A'$ such that
      $|A| = m$ and $|A'| = |B_i| - m$.
    $\mathcal{C}_{k+1} \leftarrow (\mathcal{C}_k \cup \{A, A'\}) \backslash B_i$.
    $k \leftarrow k + 1$.
end
Return $T_n, \ldots, T_2, \mathcal{C}_n, \ldots, \mathcal{C}_2$.
```

Algorithm 1: Kingman's coalescent (forward-time version).

```
Initialize $\mathcal{C}_1 = \{\{1, \ldots, n\}\}, k = 1$.
while $k < n$ do
    Sample $T_{k+1} \sim \text{Exp}(k(k+1)/2)$.
    Sample $B_i$ from $\mathcal{C}_k = \{B_1, \ldots, B_k\}$ with probability proportional
      to $(|B_i| - 1)/(n - k)$.
    Sample $m \sim \text{BetaBinomial}(|B_i|; \beta, \beta)$ conditioned on
      $1 \leq m \leq |B_i| - 1$.
    Randomly partition $B_i$ into non-empty subsets $A, A'$ such that
      $|A| = m$ and $|A'| = |B_i| - m$.
    $\mathcal{C}_{k+1} \leftarrow (\mathcal{C}_k \cup \{A, A'\}) \backslash B_i$.
    $k \leftarrow k + 1$.
end
Return $T_n, \ldots, T_2, \mathcal{C}_n, \ldots, \mathcal{C}_2$.
```

Algorithm 2: β-splitting coalescent model.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

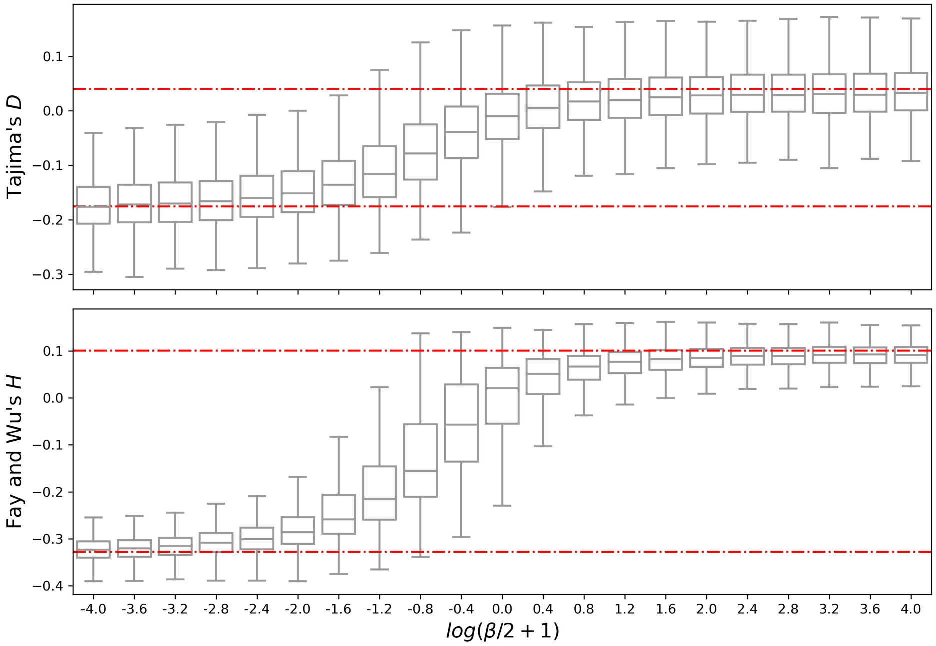


Figure 1: Empirical distributions of Tajima's D and Fay and Wu's H under different tree topologies. Going from left to right, tree structure goes from caterpillar to balanced. Red lines represent the averaged limiting cases.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

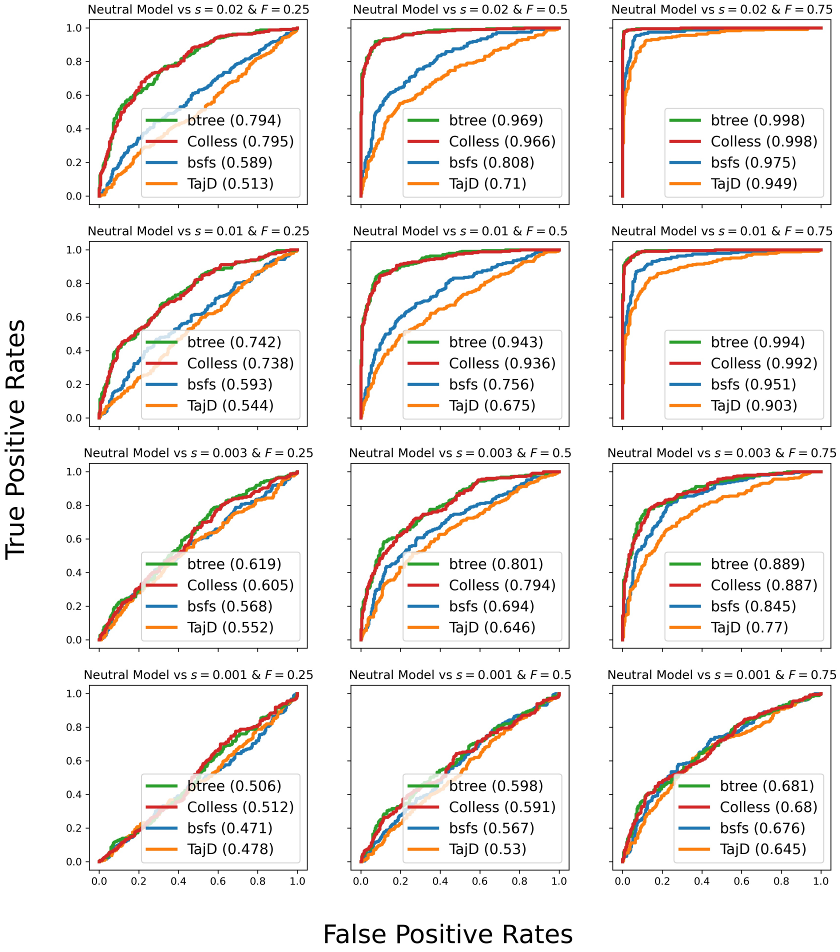


Figure 2: ROC curves for positive genic selection. s represents selective advantage of the mutation and F represents allele frequency of the mutation in the sample.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

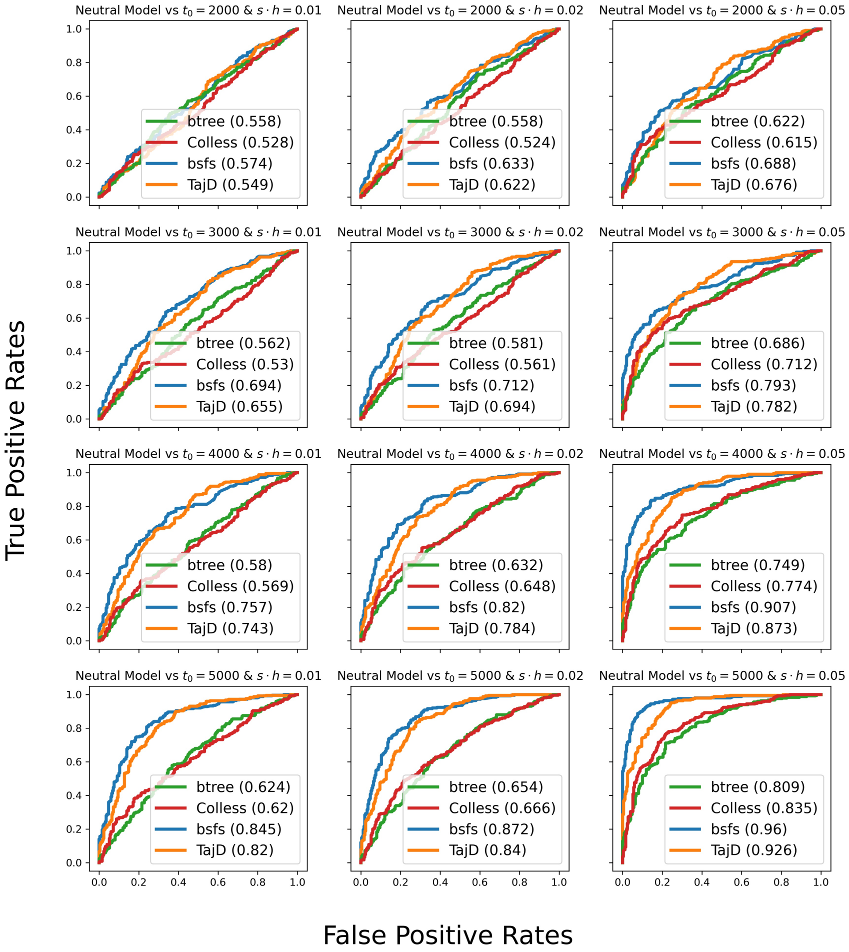
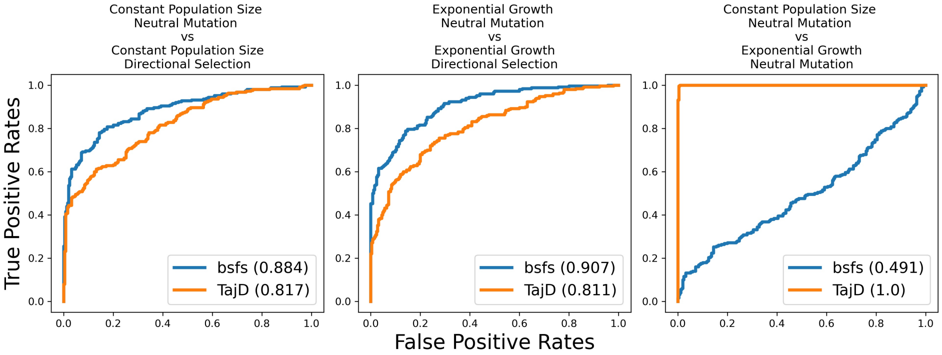
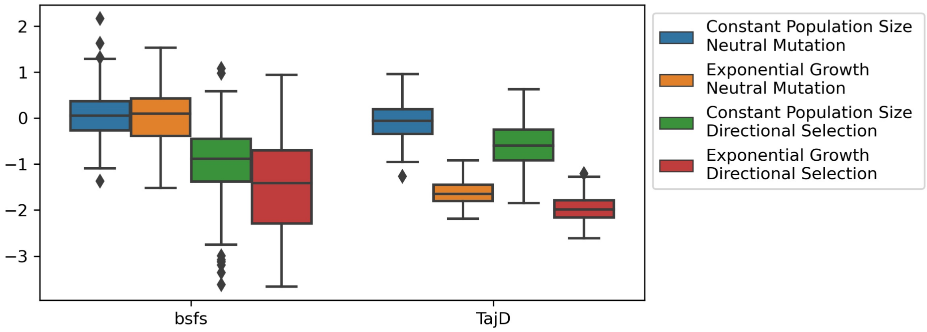


Figure 3: ROC curves for advantageous heterozygote mutation simulations. s represents selective advantage of the mutation, h is the dominance factor. t_0 represents how many generations ago the advantageous mutations were introduced into the sample.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



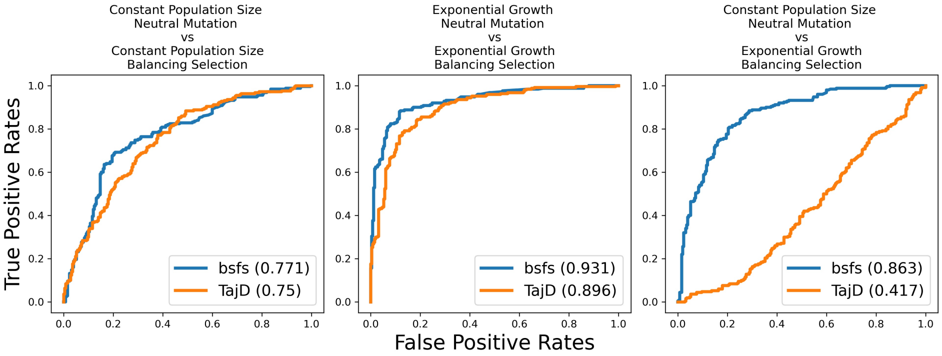
(a) ROC curves with AUC scores



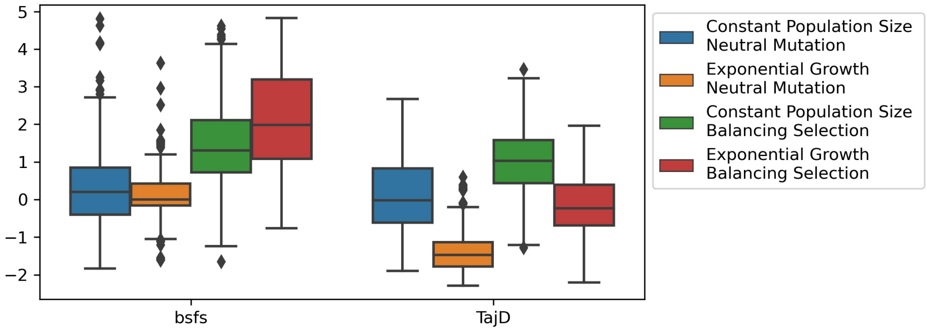
(b) Box Plot

Figure 4: Directional selection under constant population or exponentially growing size histories. bsfs is our SFS based method and TajD is Tajima's D. (a) bsfs performs better for detecting true signals in the first two figures. In the third figure, D is picking up a false positive signal with respect to detecting selection. (b) Under neutrality, bsfs has a zero centered empirical distribution regardless of the true population size history, whereas the distribution of D is shifted.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) ROC curves with AUC scores



(b) Box Plot

Figure 5: Balancing selection under constant population or exponentially growing size histories. (a) bsfs performs better for detecting true signals in the first two figures. In the third figure Tajima's D fails to detect selection. (b) Under neutrality, bsfs has a zero-centered empirical distribution and balancing selection shifts the distribution upward. Balancing selection shifts D to positive values but exponential growth pulls it downward.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

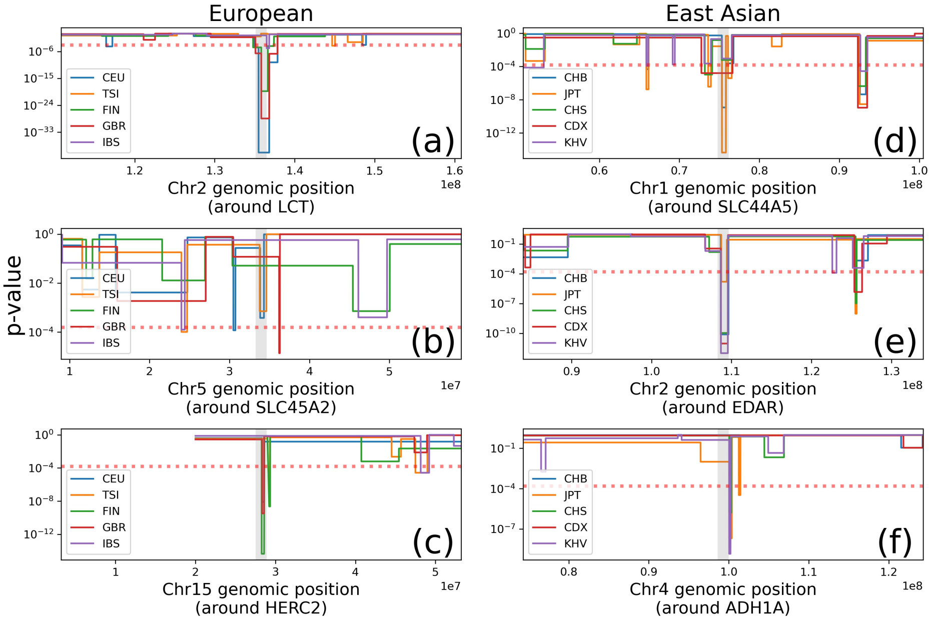


Figure 6: Results of directional selection p-value scan for 1000 Genomes Project using median centered btree (Section S1.2.1). The Bonferonni-corrected significance level is 1.6×10^{-4} (Red dashed line). Significant populations for each gene: (a) CEU, GBR, FIN; (b) None; (c) FIN, CEU, GBR; (d) JPT, CHB, CDX; (e) KHV, CDX, CHB, CHS, JPT; (f) KHV, JPT, CHS. The interval spanned by each gene is shaded in grey.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

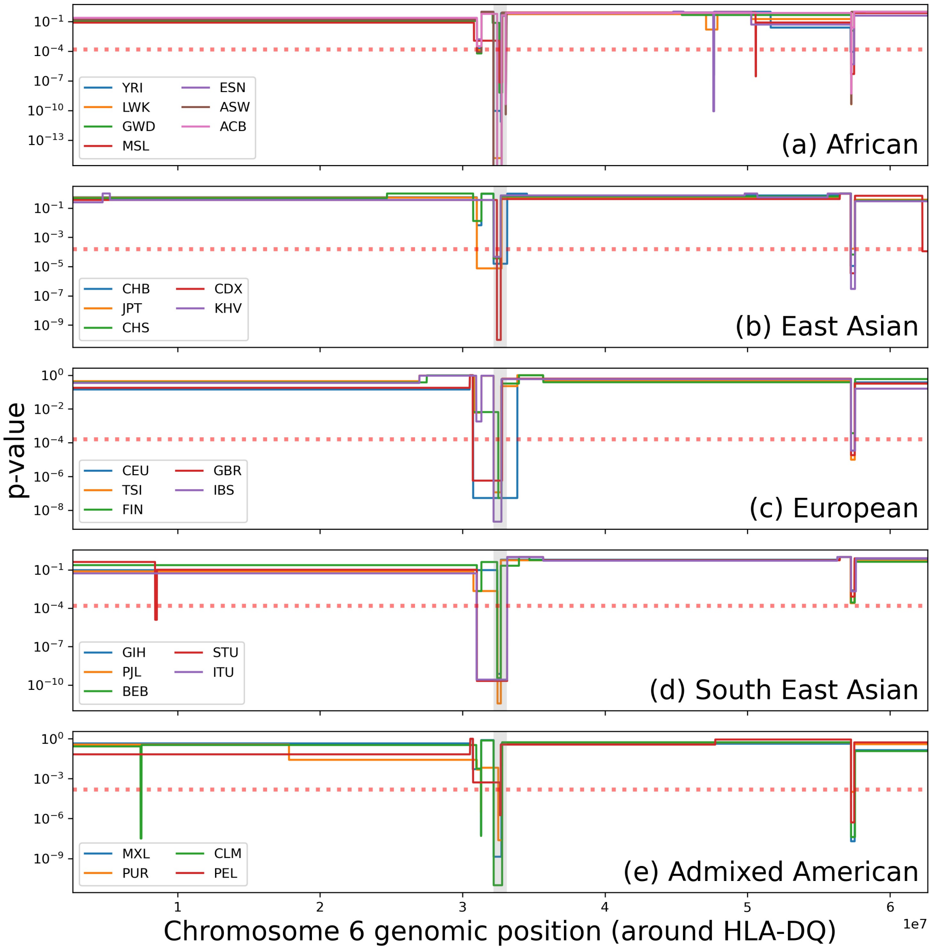


Figure 7: Genome Scan p-values of the bsfs segments around $HLA\text{-}DQ$. Most of the 1000 Genomes subpopulations have a pronounced balancing selection signal in this region.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Materials

S1 Data analysis pipeline

In the data analysis we will refer SFS-based β-splitting estimate as `bsfs` and tree-based β-splitting estimate as `btree`.

S1.1 Simulation studies

All simulations were performed using SLiMv3 (Haller and Messer, 2019). An example for the complete simulation pipeline can be seen in Figure S1.

(a) For each simulation, we have a set of parameters θ, and we run that set 250 iterations with a different seed.

(b) In SLiM we drop non-neutral mutations and run it until a stopping condition, and then record the tree-sequence.

(c) SLiM is a forward simulator and does not guarantee a common ancestor. After the simulation is completed, genealogies with multiple roots are recapitated using the procedure described by Haller et al. (2019), and neutral mutations are introduced.

(d) `bsfs` is inferred from the allele frequency spectrum. For `btree` we estimated tree-sequences from the genotype matrix using using `tsinfer` (Kelleher et al., 2019a).

(e) A simulated region covers many trees (because of the recombination). In order to represent the region, we combine these `btree`'s by taking span weighted averages. An example calculation is shown in Figure S1e. There are four `btree` estimates in this 10kb region, each spanning a different non-overlapping region. For this simulation `btree` value is calculated as:

$$\hat{\beta} = \frac{612 \times \hat{\beta}_1 + (3239 - 612) \times \hat{\beta}_2 + (8739 - 3239) \times \hat{\beta}_3 + (10001 - 8739) \times \hat{\beta}_4}{10000}$$

S1.2 Real data analysis

We tested our model on 1000 Genomes Project. A toy example of our analysis pipeline can be seen in Figure S2. For each 26 populations we repeat the following process:

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(a) The genome-wide SFS is computed for each subpopulation.

(b) Population-specific size histories η are inferred using these empirical frequency spectra.

(c) For SFS-based analysis, genomes are divided into intervals as described below, and the local frequency spectrum is calculated for each interval conditional on the size history function inferred in the previous step.

(d) Tree-sequences subdivide the genome into disjoint regions spanned by each local tree.

(e) SFS-based estimates are obtained by estimating bsfs using a sliding window along the genome.

(f) Tree-sequence estimates are performed for each local genealogy.

(g) Tree-based estimates are converted to genomic coordinates using the averaging procedure described above.

(h) To combine these spatially correlated estimates, we use a changepoint detection procedure to aggregate the bsfs and btree estimates into a set of piecewise-constant functions (Celisse et al., 2018; Truong, Oudre, and Vayatis, 2020). Heuristically, we chose a fixed number of change-points for all populations. We chose in a way that we expect a change point on the average of 10Mb (3,100Mb/10Mb = 310 change points).

(i) Finally, p-values are computed for each segment using the procedure described in the Section S1.2.3.

Genomic scan statistics depend on the choice of window size, and stride (number of base pairs between the start of each consecutive window). Since we are using estimated tree-sequences for 26 populations, the start and end points of these estimated trees are different in each population. In real data analysis we constructed two different methods of window statistics for directional selection and balancing selection. For directional selection, we are seeking population specific signal of the selection (that is the reason why we use a median-centered β estimate, Section S1.2.1). But in order to compare btree statistic with bsfs (and other neutrality tests) within the population and with other btree's across the populations we need them to be defined on the same positions on the genome. To overcome this we first estimate tree-based statistics (btree and Colless) for each tree. Second we define the windows in base pairs (we defined window size 10kbp and stride

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

size 5kbp) and calculate SFS-based statistics (bsfs and other neutrality tests) for these windows. Finally we take averages of tree-based statistic inside each windows (See Section S1.1 part (e)). In order to detect long term balancing selection, we do not need the windows to be defined in the same regions. Since we are not looking for a population specific signal. After some trial and error, we see that windows defined by estimated tree locations give better results for balancing selection. Instead of sliding base pairs, we chose window size of 64 trees, and a stride of 32 trees, we calculated the SFS-based statistics on those regions.

In directional selection setup, for bsfs, this window-size and stride setup resulted 520,940 windows along the human genome for bsfs. This required us to solve $520940 \times 26 = 13,544,440$ optimization problems. For btree, each population has different number of trees, and in total it resulted 142,637,760 optimization tasks. Together with other statistics, all calculations required less than four hours on a cluster. For the balancing selection setup, number of tree estimates doesn't change so we do not need to estimate btree again. But we estimated bsfs again for the different windows which required solving an additional $8,914,860$ optimization problems and this finished approximately in 2 hours.

S1.2.1 Median-centered estimates of β

Some regions on the chromosome experience the similar evolutionary history among all human populations. For some regions, this causes a pronounced spike in $|\beta|$ for all 26 populations in 1000 Genomes Project data, confounding our ability to detect population-specific signals of selection. To correct for this, we performed median-centering for windowed statistic: let $\mathcal{W}$ defines the set of windows along the chromosome, and $\hat{\beta}^0_{w,p}$ be the β estimate of window $w \in \mathcal{W}$ for each of the 26 1kg subpopulations

$$p \in \mathcal{P} := \{\mathrm{CEU}, \mathrm{CHB}, \ldots, \mathrm{YRI}\}.$$

Then the median-centered estimate of $\hat{\beta}_{w,p}$ is defined as

$$\hat{\beta}_{w,p} := \hat{\beta}^0_{w,p} - \mathrm{median}\left\{\hat{\beta}_{w,x} : x \in \mathcal{P}\right\}.$$

S1.2.2 Combining multiple $\hat{\beta}$ in each segments

Each segment decided by change point detection spans multiple windows. We average these windowed estimates of β and get a single estimate that represents each segment. To combine window estimates we used weighted

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

averaging. Given n_g windows spanning a given segment, we define $\bar{\beta}_{\boldsymbol{\omega}} = \sum_{i=1}^{n_g} \omega_i \hat{\beta}_i$ to be a weighted sum of estimated $\hat{\beta}$ parameters, where $\boldsymbol{\omega}$ is a weight vector. After some experimenting, we found that choosing the entries of $\boldsymbol{\omega}$ to be proportional to the proximity of the mid-point of the segment worked well, and all reported results are based on that choice of weights.

S1.2.3 Significance testing

Since the $\hat{\beta}$ are maximum likelihood estimates, $\bar{\beta}_{\boldsymbol{\omega}}$ has approximately a normal distribution. To form p-values we therefore require the mean and variance of this statistic under the null hypothesis. Under neutrality, $\mathbb{E}\bar{\beta}_{\mathbf{w}} = \mu_g$, where $\mu_g \approx 0$ is the chromosome-wide average which is determined empirically. To calculate $\text{var}(\bar{\beta}_{\boldsymbol{\omega}})$ we need to consider the dependence between sequential estimates, which is non-zero due to linkage. We define

$$\sigma_g^2 = \text{var}\left(\sum_{i=1}^{n_g} \omega_i \hat{\beta}_i\right) = C_0 \sum_{i=1}^{n_g} \omega_i^2 + 2C_1 \sum_{i=1}^{n_g-1} \omega_i \omega_{i+1} + \cdots + 2C_{n_g-1} \sum_{i=1}^{1} \omega_i \omega_{i+n_g-1} \tag{20}$$

where $C_i = \text{cov}(\hat{\beta}_j, \hat{\beta}_{j+i})$ is the lag-i autocovariance term, which is assumed to be stationary (does not depend on j). The coefficients C_i were are estimated empirically using chromosome-wide averages.

For each location we perform a one-sided test to determine whether $\bar{\beta}_{\mathbf{w}}$ is abnormally high (signifying balancing selection) or low (directional selection). The corresponding p-values are

$$\begin{aligned} p^+ &= 1 - \Phi\left(\frac{\bar{\beta}_{\mathbf{w}} - \mu_g}{\sigma_g}\right) \\ p^- &= \Phi\left(\frac{\bar{\beta}_{\mathbf{w}} - \mu_g}{\sigma_g}\right) \end{aligned} \tag{21}$$

Finally, the p-values are Bonferroni corrected to account for multiple testing. In this case for 1000 genomes project population, we defined 311 segments, then the significance level will be equal to $0.05/311 \approx 1.6 \times 10^{-4}$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S2 Supplemental algorithms

```
Initialize $\mathcal{C}_n = \{\{1\}, \dots, \{n\}\}, k = n$.
while $k > 1$ do
    Sample $T_k \sim \mathrm{Exp}(k(k-1)/2)$.
    Sample $B, B' \in \mathcal{C}_k$ uniformly without replacement.
    $\mathcal{C}_{k-1} \leftarrow (\mathcal{C}_k \cup \{B \cup B'\}) \backslash B \backslash B'$.
    $k \leftarrow k - 1$.
end
Return $\{(T_n, \dots, T_2), (\mathcal{C}_n, \dots, \mathcal{C}_2)\}$.
```

Algorithm S1: Kingman's coalescent.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S3 Simulation settings and code

Scenario	$N_e(0)$	g	t_g	s	h
1	2×10^3	0	–	0	0.5
2	10^3	0.01	250	0	0.5
3	2×10^3	0	–	0.05	0.5
4	10^3	0.01	250	0.05	0.5

Table S1: Simulation settings for directional selection experiment. $N_e(0)$ is population size at the start of the simulation, g is the growth rate of exponential growth, t_g is the generations ago when exponential growth starts prior to sampling, s is the selective coefficient of the beneficial mutation and h is the dominance coefficient.

Scenario	$N_e(0)$	μ_s	t_m	g	t_g	s	h
1	2×10^3	0	–	0	–	0	0.5
2	10^3	0	–	0.01	250	0	0.5
3	2×10^3	10^{-8}	4250	0	–	0.002	25
4	10^3	10^{-8}	4250	0.01	250	0.002	25

Table S2: Simulation settings for balancing selection experiment. μ_s is the mutation rate for advantageous mutation, t_m is the number of generations age that selection began, and the rest of the parameters are as in Table S1.

S3.1 Directional selection

```
initialize() {
    if (exists("slimgui")) {
        defineConstant("simID", 1);
    }

    defineConstant("pmut", asInteger([L]/2)); // Mutation
       position
    initializeTreeSeq();
    initializeMutationRate([mu]*0.08);
    initializeMutationType("m1", 0.5, "f", 0.0);
```

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

```
    initializeMutationType("m3", [h], "f", [s]);
    initializeGenomicElementType("g1", m3, 1);
    initializeGenomicElement(g1, 0, asInteger([L]));
    initializeRecombinationRate([r]);
}
1 {
    sim.addSubpop("p1", asInteger([Ne]));
}

1 late(){
    defineGlobal("bar", 0);
    //target = sample(p1.genomes, 1);
    //target.addNewDrawnMutation(m3, pmut);
    sim.treeSeqOutput("/scratch/stats_dept_root/stats_dept/enes
        /slim_" + simID + ".trees");
}

1:[Until] late() {
    if ([start]<sim.generation){
        newSize = asInteger(round([rep]^(sim.generation-[start
            ]) * [Ne]));
        p1.setSubpopulationSize(newSize);
    }

    m3muts = sim.mutationsOfType(m3);
    freqs = sum(sim.mutationFrequencies(NULL, m3muts));
    if ([reset_lost]){
        if (sim.countOfMutationsOfType(m3) < bar){
            cat(simID + ": RESTARTING");
            sim.readFromPopulationFile("/scratch/
                stats_dept_root/stats_dept/enes/slim_" + simID
                + ".trees");
            setSeed(rdunif(1, 0, asInteger(2^62) - 1));
            defineGlobal("bar", 0);
        }
    }
}
[Until] late(){
    m3muts = sim.mutationsOfType(m3);
    freqs = sum(sim.mutationFrequencies(NULL, m3muts));
    print(freqs);
    sim.treeSeqOutput("trees/"+simID+".trees");
    sim.simulationFinished();
}
```

S3.2 Balancing selection

```
initialize() {
    if (exists("slimgui")) {
```

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

```
        defineConstant("simID", 1);
    }

    defineConstant("pmut", asInteger([L]/2)); // Mutation
       position
    initializeTreeSeq();
    initializeMutationRate([mu]*0.08);
    initializeMutationType("m1", 0.5, "f", 0.0);
    initializeMutationType("m3", [h], "f", [s]);
    initializeGenomicElementType("g1", m3, 1);
    initializeGenomicElement(g1, 0, asInteger([L]));
    initializeRecombinationRate([r]);
}
1 {
    sim.addSubpop("p1", asInteger([Ne]));
}

1 late(){
    defineGlobal("bar", 0);
    sim.treeSeqOutput("/scratch/stats_dept_root/stats_dept/enes
       /slim_" + simID + ".trees");
}

1:[Until] late() {
    if ([start]<sim.generation){
        newSize = asInteger(round([rep]^(sim.generation-[start
           ]) * [Ne]));
        p1.setSubpopulationSize(newSize);
    }

    m3muts = sim.mutationsOfType(m3);
    freqs = sum(sim.mutationFrequencies(NULL, m3muts));
    if ([reset_lost]){
        if (sim.countOfMutationsOfType(m3) < bar){
            cat(simID + ": RESTARTING");
            sim.readFromPopulationFile("/scratch/
               stats_dept_root/stats_dept/enes/slim_" + simID
               + ".trees");
            setSeed(rdunif(1, 0, asInteger(2^62) - 1));
            defineGlobal("bar", 0);
        }
    }
}
[Until] late(){
    m3muts = sim.mutationsOfType(m3);
    freqs = sum(sim.mutationFrequencies(NULL, m3muts));
    print(freqs);
    sim.treeSeqOutput("trees/"+simID+".trees");
    sim.simulationFinished();
```

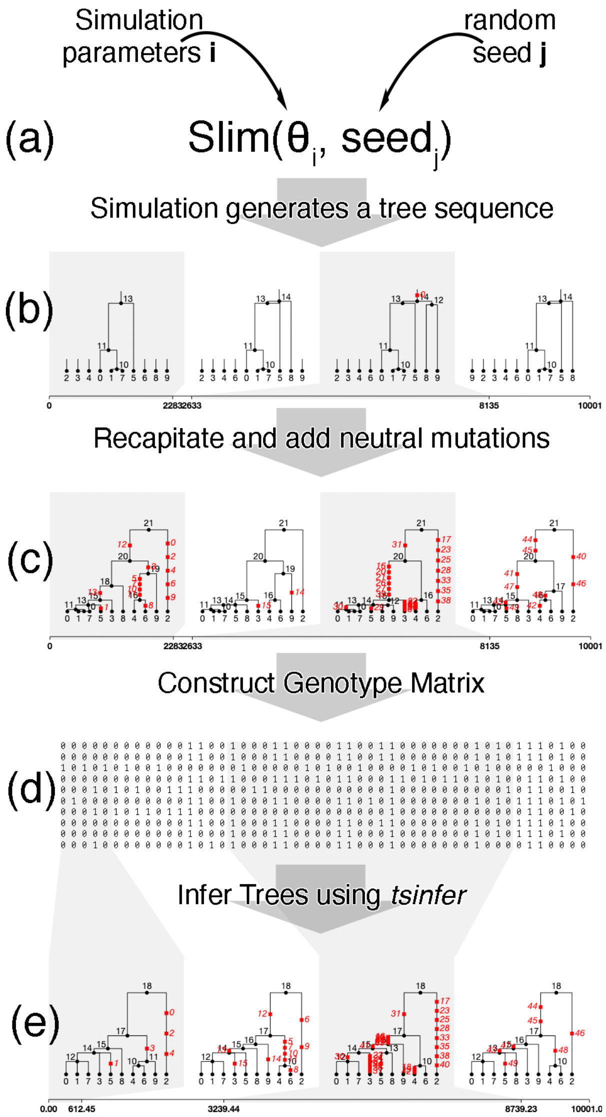
bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

```
}
```

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S4 Supplemental figures

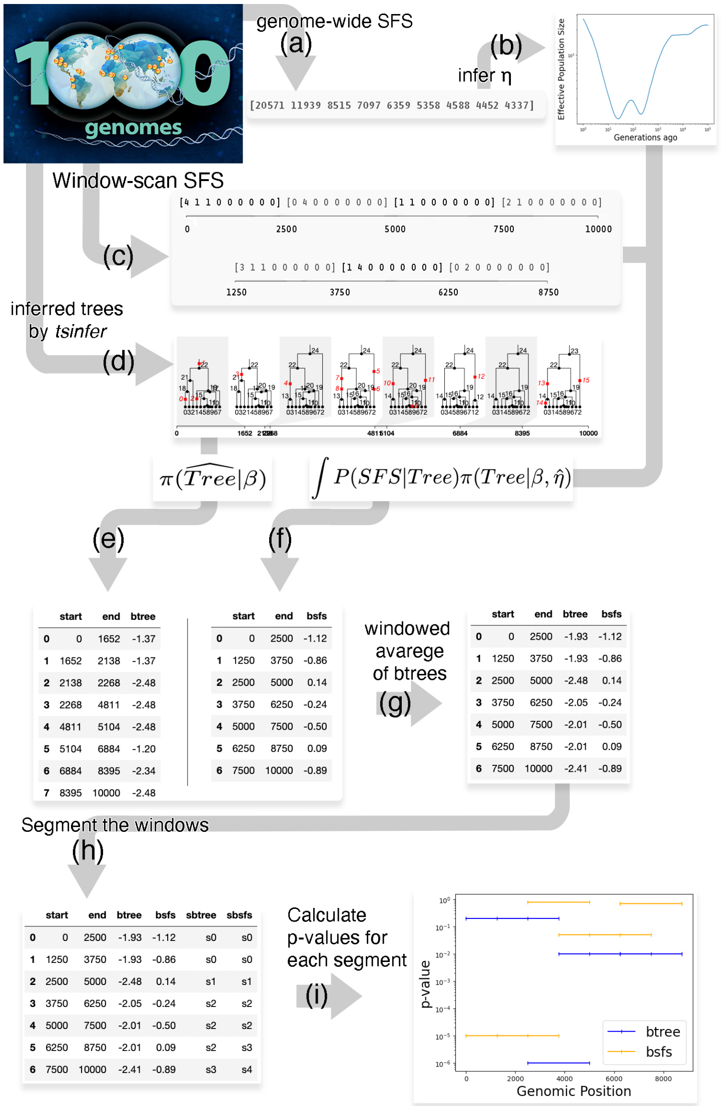
bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


Simulation parameters **i**
random seed **j**
(a)
Slim(θ_i, seed_j)
Simulation generates a tree sequence
(b)
0
22832633
8135
10001
Recapitate and add neutral mutations
(c)
0
22832633
8135
10001
Construct Genotype Matrix
(d)
Infer Trees using *tsinfer*
(e)
0.00
612.45
3239.44
8739.23
10001.0

```
0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 1 1 0 0 0 0 1 1 0 0 1 1 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 1 1 0 0 0 0 1 1 0 0 1 1 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 1 0 0
1 0 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 1 0 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 0 1 0 1 1 0 0 0 1 1 0 1 0 1 1 0 1 0 1 0 0 1 1 0 1 0 0
0 0 0 1 0 1 0 1 0 0 1 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 1 0
0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 1 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 1 0 1
0 0 0 1 0 1 0 1 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0
0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 1 1 0 0 0 0 1 1 0 0 1 1 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 1 0 0
0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0
```

Figure S1: Simulation pipeline. (a) In each set of simulations we fix a group of parameters (θ) and a random seed. (b) After simulation is finished, we save the tree sequence of the true genealogy. It can be seen that trees have multiple roots and the sample only have one mutation at the third tree. (c) We recapitate the trees and add the neutral mutations. (d) Genotype matrix is extracted from the tree sequence and used to estimate bsfs. (e) Tree sequences are inferred from genotype matrices and used to estimate btree.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



```
[20571 11939 8515 7097 6359 5358 4588 4452 4337]
```

```
[4 1 1 0 0 0 0 0 0] [0 4 0 0 0 0 0 0 0] [1 1 0 0 0 0 0 0 0] [2 1 0 0 0 0 0 0 0]
0          2500          5000          7500          10000

[3 1 1 0 0 0 0 0 0] [1 4 0 0 0 0 0 0 0] [0 2 0 0 0 0 0 0 0]
1250          3750          6250          8750
```

	start	end	btree
0	0	1652	-1.37
1	1652	2138	-1.37
2	2138	2268	-2.48
3	2268	4811	-2.48
4	4811	5104	-2.48
5	5104	6884	-1.20
6	6884	8395	-2.34
7	8395	10000	-2.48

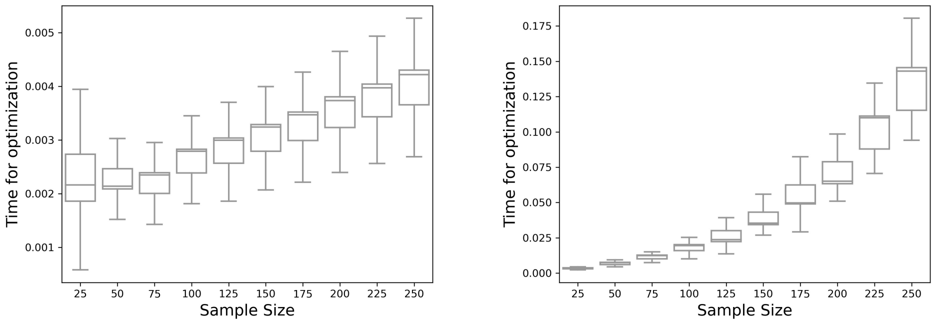
	start	end	bsfs
0	0	2500	-1.12
1	1250	3750	-0.86
2	2500	5000	0.14
3	3750	6250	-0.24
4	5000	7500	-0.50
5	6250	8750	0.09
6	7500	10000	-0.89

	start	end	btree	bsfs
0	0	2500	-1.93	-1.12
1	1250	3750	-1.93	-0.86
2	2500	5000	-2.48	0.14
3	3750	6250	-2.05	-0.24
4	5000	7500	-2.01	-0.50
5	6250	8750	-2.01	0.09
6	7500	10000	-2.41	-0.89

	start	end	btree	bsfs	sbtree	sbsfs
0	0	2500	-1.93	-1.12	s0	s0
1	1250	3750	-1.93	-0.86	s0	s0
2	2500	5000	-2.48	0.14	s1	s1
3	3750	6250	-2.05	-0.24	s2	s2
4	5000	7500	-2.01	-0.50	s2	s2
5	6250	8750	-2.01	0.09	s2	s3
6	7500	10000	-2.41	-0.89	s3	s4

Figure S2: Real data analysis pipeline. (a) We calculate genome-wide SFS for the sample. (b) Using this we infer population size histories (η). (c) We calculate windowed statistic of SFS. Window size is 2500 and stride is 1250. With the region of size 10000, we get 7 windows. (d) From genotype matrix, we infer trees using `tsinfer`. (d) For each tree in the tree sequence, we estimate btree. (f) For each windowed-statistic SFS we estimate bsfs. (g) We take average btrees inside bsfs regions by taking span weighted avarage of each. (h) We apply the change point detection method to define segments. (i) We calculate the p-values for each segment.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) β from tree-sequences

(b) β from SFS

Figure S3: Time-complexity of estimating β for a single optimization task.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

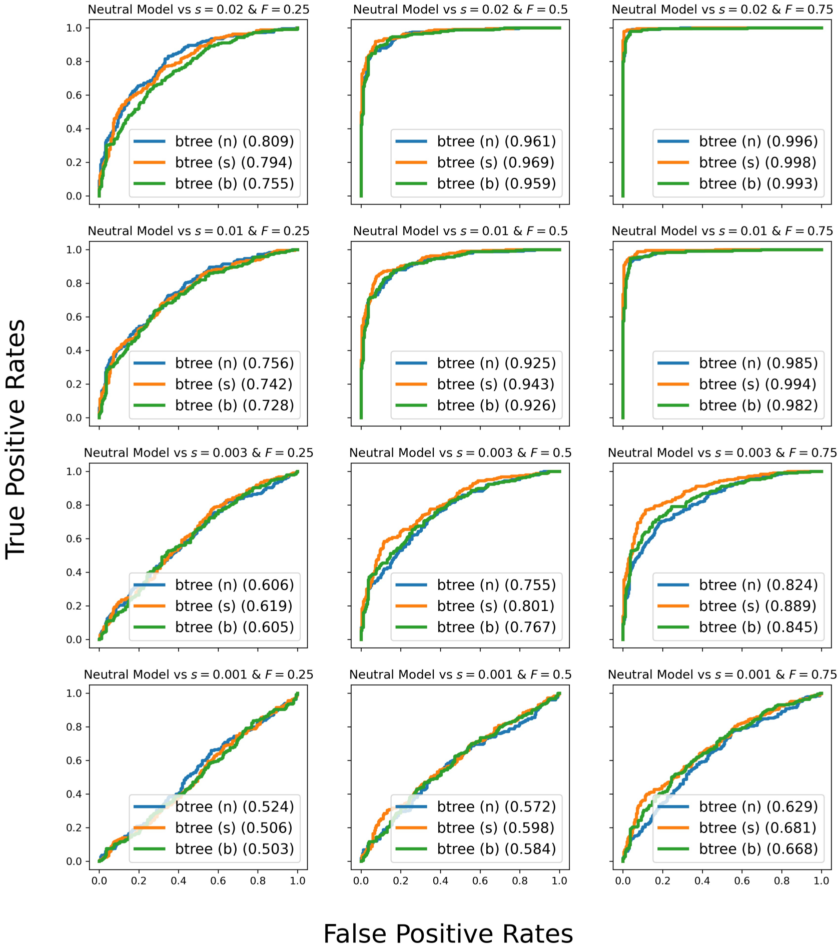


Figure S4: ROC curves for positive genic selection using btree with three different choices of likelihood weights. s represents selective advantage of the mutation and F represents allele frequency of the mutation in the sample. Letter inside the parenthesis represents the likelihood weighting method; (n):no weighting, (s): size of the internal node, (b): total amount of branch length

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

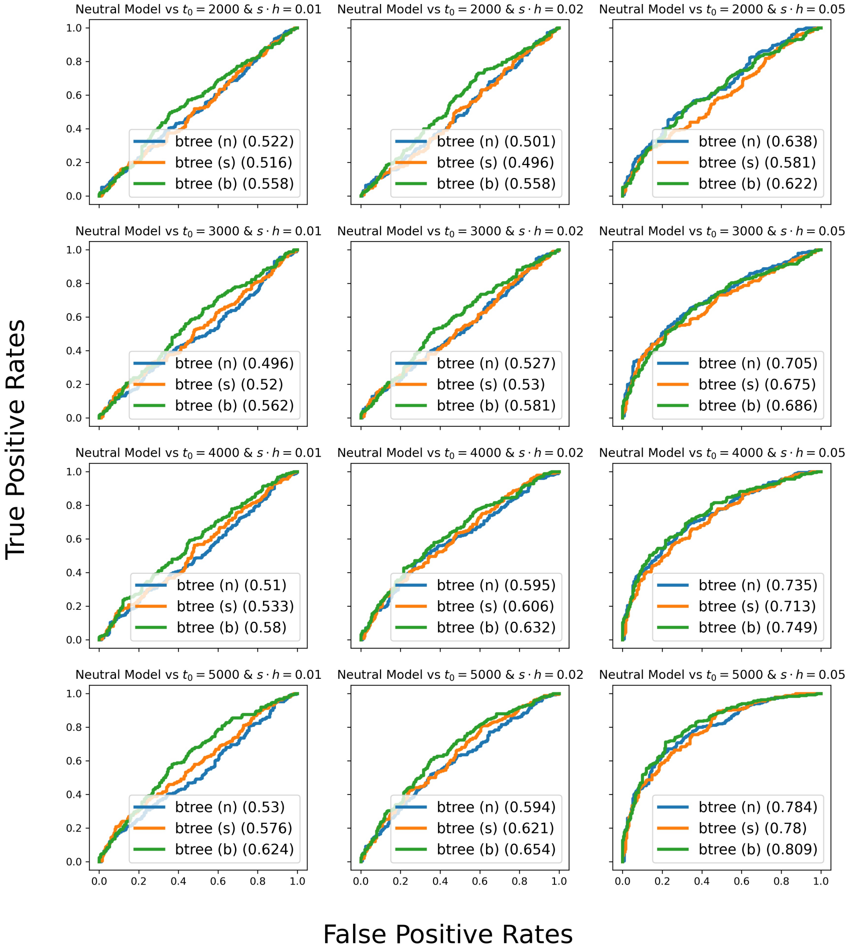


Figure S5: ROC curves for advantageous heterozygote mutation simulations using btree with three different choices of likelihood weights. s represents selective advantage of the mutation, h is the dominance factor. t_0 represents how many generations ago the advantageous mutations were introduced into the sample. Letter inside the parenthesis represents the likelihood weighting method; (n):no weighting, (s): size of the internal node, (b): total amount of branch length

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

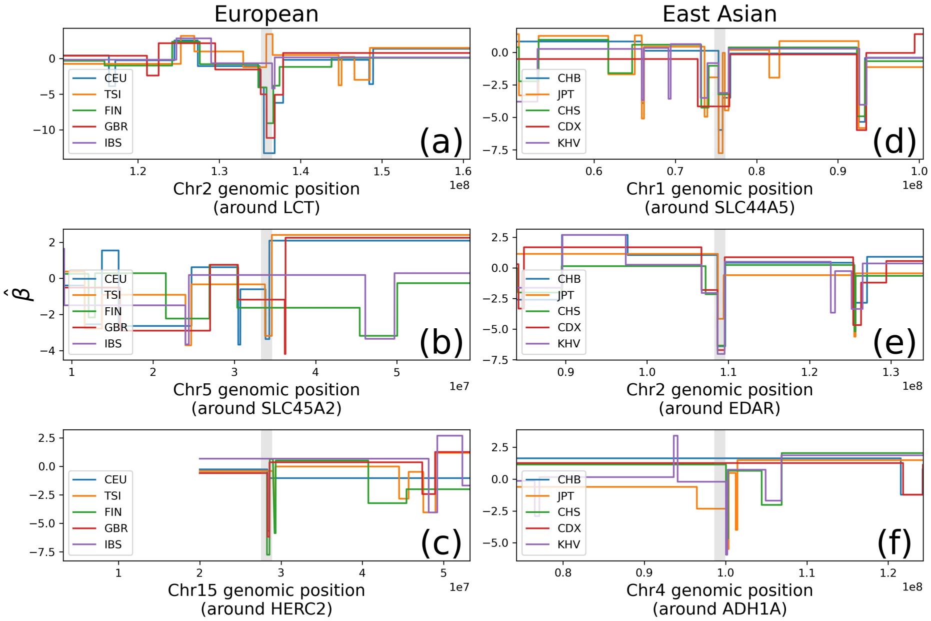


Figure S6: Directional selection examples for 1000 Genomes Project. Segmented genome-scan β estimates by median centered btree are provided. See Figure 6

.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

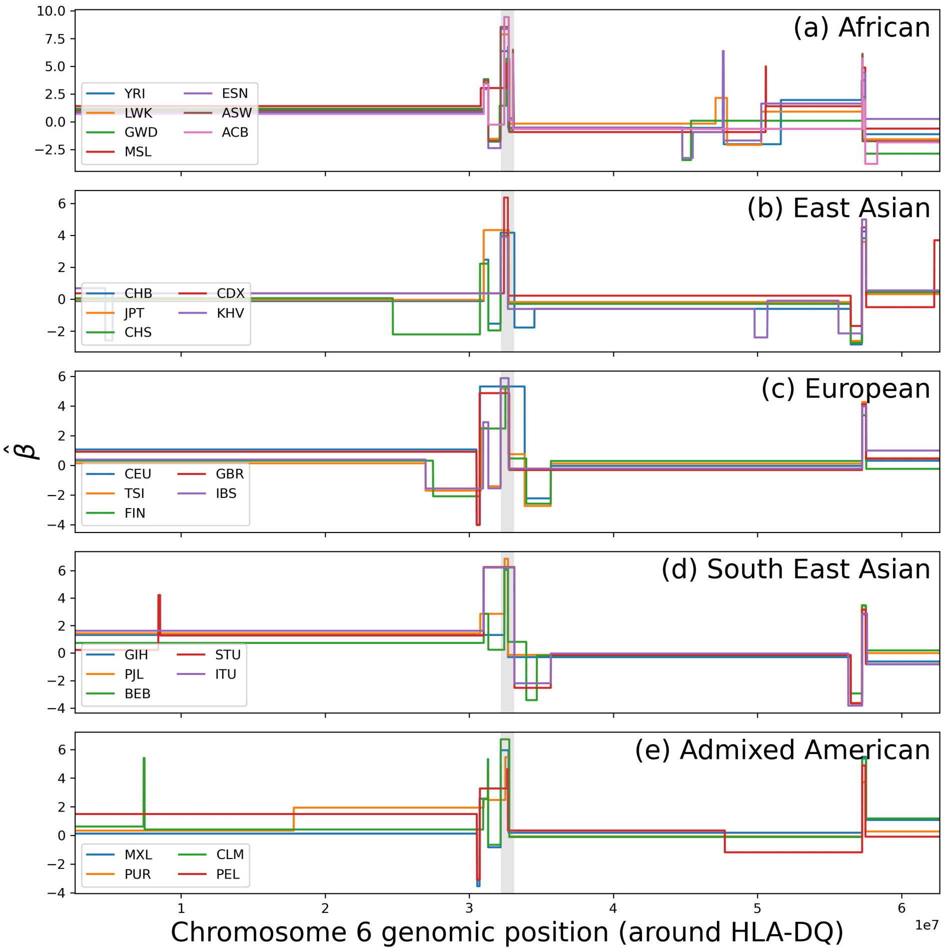


Figure S7: Genome Scan β estimates by bsfs segments around *HLA-DQ*. See Figure 7.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S5 Supplemental Tables

Table S3: Most significant directional selection hits for East Asian

Population	Chromosome	start	end	p value	Average $\hat{\beta}$	genes
JPT	1	0.75	0.76	4×10^{-15}	-2.27	ACADM, RABGGTB, MSH4
KHV	2	0.17	0.18	4×10^{-13}	-2.39	RAD51AP2, VSNL1
KHV	2	1.09	1.10	1×10^{-12}	-2.02	RANBP2, CCDC138, EDAR, SH3RF3, SEPTIN10
CHB	15	0.64	0.65	1×10^{-12}	-1.79	CIAO2A, SNX1, SNX22, PPIB, CSNK1G1, PCLAF, TRIP4, ZNF609, OAZ2, RBPMS2, PIF1, PLEKHO2, ANKDD1A, SPG21, MTFMT, SLC51B, RASL12, KBTBD13, UBAP1L
CDX	2	0.43	0.44	4×10^{-10}	-1.89	PLEKHH2, C1GALT1C1L, DYNC2LI1, ABCG5, ABCG8, LRPPRC
JPT	4	1.43	1.44	7×10^{-10}	-1.57	SMARCA5, FREM3, GYPE, GYPB, GYPA
CDX	1	0.92	0.93	1×10^{-09}	-1.39	GFI1, EVI5, RPL5, DIPK1A, MTF2, TMED5, CCDC18, DR1, FNBP1L
JPT	3	0.17	0.18	1×10^{-09}	-1.15	PLCL2, TBC1D5, SATB1
KHV	4	1.00	1.00	1×10^{-09}	-2.42	DDIT4L
CDX	2	1.97	1.98	2×10^{-09}	-1.91	SF3B1, COQ10B, HSPD1, HSPE1, MOB4, RFTN2

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S4: Most significant directional selection hits for European

Population	Chromosome	start	end	p value	Average $\hat{\beta}$	genes
CEU	2	1.35	1.37	2×10^{-40}	-2.31	R3HDM1, UBXN4, LCT, MCM6, DARS1, CXCR4, THSD7B
GBR	12	1.11	1.13	3×10^{-18}	-1.50	ATXN2, BRAP, ACAD10, ALDH2, MAPKAPK5, TMEM116, ERP29, NAA25, TRAFD1, RPL6, PTPN11, RPH3A
FIN	15	0.28	0.29	4×10^{-15}	-1.73	GOLGA8F, GOLGA8G
TSI	11	0.89	0.89	5×10^{-13}	-1.63	TYR
TSI	5	1.30	1.31	4×10^{-11}	-1.33	CHSY3, HINT1, LYRM7, CDC42SE2
GBR	8	0.12	0.13	5×10^{-11}	-1.98	DEFB130A, FAM86B2
FIN	6	0.28	0.29	9×10^{-11}	-1.35	GPX5, ZBED9
GBR	15	0.75	0.75	3×10^{-10}	-1.39	CLK3, EDC3, CYP1A1, CYP1A2, CSK, LMAN1L, CPLX3, ULK3, SCAMP2, MPI, FAM219B, COX5A, RPP25, SCAMP5, PPCDC
GBR	6	0.35	0.35	8×10^{-10}	-1.42	TCP11, SCUBE3, ZNF76, DEF6, PPARD
CEU	1	1.00	1.00	1×10^{-09}	-1.87	AGL

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S5: Most significant directional selection hits for African

Population	Chromosome	start	end	p value	Average $\hat{\beta}$	genes
ESN	19	0.43	0.44	4×10^{-17}	-2.19	PSG5, PSG4, PSG9, CD177, TEX101, LYPD3, PHLDB3, ETHE1, ZNF575, XRCC1
ESN	6	0.53	0.53	5×10^{-16}	-2.74	TMEM14A
GWD	6	0.59	0.62	9×10^{-14}	-2.25	KHDRBS2
LWK	8	0.44	0.47	1×10^{-12}	-1.60	SPIDR
ESN	11	0.49	0.49	2×10^{-11}	-1.60	TRIM49B, TRIM64C
YRI	1	1.62	1.62	2×10^{-11}	-1.80	DUSP12, ATF6
ESN	1	1.54	1.54	7×10^{-11}	-1.73	CRTC2, SLC39A1, CREB3L4, JTB, RAB13, RPS27, NUP210L
ESN	3	1.25	1.26	2×10^{-09}	-1.61	SNX4, OSBPL11
ASW	13	0.52	0.52	4×10^{-09}	-1.43	NEK3
YRI	19	0.39	0.39	4×10^{-09}	-1.68	LGALS7, LGALS7B, LGALS4, ECH1, HNRNPL, RINL, SIRT2

Table S6: Most significant directional selection hits for Ad-Mixed American

Population	Chromosome	start	end	p value	Average $\hat{\beta}$	genes
PUR	22	0.29	0.29	1×10^{-07}	-1.87	EMID1, RHBDD3, EWSR1, GAS2L1, RASL10A, AP1B1, RFPL1
CLM	6	1.35	1.35	2×10^{-07}	-2.18	AHI1
CLM	8	0.44	0.48	2×10^{-07}	-1.54	SPIDR
PUR	3	1.29	1.30	4×10^{-07}	-1.91	EFCAB12, MBD4, IFT122, RHO, H1-8, PLXND1
PUR	17	0.37	0.37	5×10^{-07}	-2.39	ACACA
MXL	2	0.24	0.25	8×10^{-07}	-1.99	NCOA1
PUR	12	0.44	0.48	1×10^{-06}	-0.94	NELL2, DBX2, ANO6, ARID2, SCAF11, SLC38A1, SLC38A2, SLC38A4, AMIGO2, PCED1B
MXL	8	0.86	0.87	1×10^{-06}	-1.71	PSKH2, ATP6V0D2, SLC7A13, WWP1, RMDN1, CPNE3, CNGB3, CNBD1
PUR	11	0.13	0.13	1×10^{-06}	-1.92	TEAD1
PUR	7	1.29	1.29	2×10^{-06}	-1.84	ATP6V1F, ATP6V1FNB, IRF5, TNPO3

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S7: Most significant directional selection hits for South East Asian

Population	Chromosome	start	end	p value	Average $\hat{\beta}$	genes
BEB	16	0.34	0.47	3×10^{-17}	-1.45	SHCBP1, VPS35, ORC6, MYLK3, C16orf87, GPT2
GIH	15	0.48	0.49	3×10^{-12}	-1.69	DUT, FBN1
STU	17	0.44	0.44	3×10^{-12}	-1.77	HROB, ASB16, TMUB2, ATXN7L3, UBTF, SLC4A1, RUNDC3A, SLC25A39, GRN, FAM171A2
GIH	3	0.89	0.94	8×10^{-12}	-0.84	EPHA3, PROS1
GIH	2	0.97	0.97	1×10^{-11}	-1.73	FAHD2B
STU	16	0.30	0.31	4×10^{-11}	-1.23	SEPHS2, ITGAL, ZNF768, ZNF747, ZNF764, ZNF688, ZNF785, ZNF689, PRR14, FBRS, SRCAP, TMEM265, PHKG2, CCDC189, RNF40, ZNF629, BCL7C, CTF1, FBXL19, ORAI3, SETD1A, HSD3B7, STX1B, STX4, ZNF668, ZNF646, PRSS53, VKORC1, BCKDK, KAT8, PRSS8, PRSS36, FUS, PYCARD, TRIM72, PYDC1, ITGAM, ITGAX, ITGAD, COX6A2, ZNF843, ARMC5, TGFB1I1
PJL	2	2.23	2.23	5×10^{-11}	-1.60	FARSB
PJL	2	0.92	0.95	6×10^{-11}	-1.63	TEKT4, MAL, MRPS5, ZNF514, ZNF2, PROM2, KCNIP3
STU	10	0.93	0.94	1×10^{-10}	-1.20	CEP55, FFAR4, RBP4, PDE6C, FRA10AC1, LGI1, SLC35G1, PLCE1
STU	19	0.36	0.36	4×10^{-10}	-1.90	ZFP82

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S8: Most significant directional selection hits for 1000 Genomes Project Populations

Population	Chromosome	start	end	p value	Average $\hat{\beta}$	genes
CEU	2	1.35	1.37	2×10^{-40}	-2.31	R3HDM1, UBXN4, LCT, MCM6, DARS1, CXCR4, THSD7B
GBR	12	1.11	1.13	3×10^{-18}	-1.50	ATXN2, BRAP, ACAD10, ALDH2, MAPKAPK5, TMEM116, ERP29, NAA25, TRAFD1, RPL6, PTPN11, RPH3A
BEB	16	0.34	0.47	3×10^{-17}	-1.45	SHCBP1, VPS35, ORC6, MYLK3, C16orf87, GPT2
ESN	19	0.43	0.44	4×10^{-17}	-2.19	PSG5, PSG4, PSG9, CD177, TEX101, LYPD3, PHLDB3, ETHE1, ZNF575, XRCC1
ESN	6	0.53	0.53	5×10^{-16}	-2.74	TMEM14A
JPT	1	0.75	0.76	4×10^{-15}	-2.27	ACADM, RABGGTB, MSH4
FIN	15	0.28	0.29	4×10^{-15}	-1.73	GOLGA8F, GOLGA8G
GWD	6	0.59	0.62	9×10^{-14}	-2.25	KHDRBS2
KHV	2	0.17	0.18	4×10^{-13}	-2.39	RAD51AP2, VSNL1
TSI	11	0.89	0.89	5×10^{-13}	-1.63	TYR

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.12.443797; this version posted May 13, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S9: Most significant balancing selection hits for 1000 Genomes Project Popultions

Population	Chromosome	start	end	p value	Average $\hat{\beta}$	genes
ASW	6	0.322	0.327	$< 10^{-38}$	1.33	RNF5, AGER, PBX2, GPSM3, NOTCH4, TSBP1, BTNL2, HLA-DRA, HLA-DRB5, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DQA2
YRI	4	0.989	0.991	$< 10^{-38}$	4.71	METAP1
IBS	1	1.529	1.532	$< 10^{-38}$	6.43	SPRR4, SPRR1A, SPRR3, SPRR1B, SPRR2D, SPRR2A, SPRR2B, SPRR2E, SPRR2F, SPRR2G
ESN	4	0.854	0.855	$< 10^{-38}$	5.78	ARHGAP24
ASW	5	1.355	1.355	$< 10^{-38}$	6.61	NEUROG1
ACB	11	0.051	0.051	$< 10^{-38}$	5.16	OR52E1
ASW	6	0.330	0.331	$< 10^{-38}$	4.98	HLA-DPA1, HLA-DPB1
IBS	13	0.368	0.368	$< 10^{-38}$	6.59	RFXAP
STU	1	2.315	2.315	$< 10^{-38}$	5.87	TSNAX
STU	20	0.235	0.236	$< 10^{-38}$	5.76	CST9L