OXFORD GENETICS

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

Enes Dilber[1] and Jonathan Terhorst[1,*]

[1]Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, USA

*Corresponding author: jonth@umich.edu

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 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 tests in a principled way.

In this paper we derive new likelihood-based methods for detecting natural selection, which are robust to 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.

Keywords: neutrality test; site frequency spectrum; natural selection

Introduction

Understanding how species adapt to their surroundings has been a defining challenge in biology for 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 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 just a snapshot of genetic variation taken at a single point in time. Theoretical models are essential in order to decipher these convoluted signals in a principled manner.

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 *et al.* 1988). These distortions are then manifested in data as altered patterns of genetic variation. By fitting a statistical model of this process, we can learn about natural selection using observed polymorphisms.

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,b), 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 the tractability of neutral coalescent; the resulting estimators are fast, model-based, and easy to understand and implement. An important feature of our method is that 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).

Related work

We lack space to survey the panoply of methods that have been developed to study natural selection using genomic data; see recent reviews by Vitti *et al.* (2013) and Stern and Nielsen (2019). We focus here on several 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 above-mentioned 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 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 (One standard practice is to subtract the genome-wide 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)). 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 related line of work aims to detect deviations of certain topological moments from the neutral expectation. Li (2011) is a simulation-based approach for detecting natural selection by examining imbalance in the basal (top-most) split in a reconstructed coalescent tree. Li and Wiehe (2013) utilizes a moment-matched normal approximation to the distribution of a tree imbalance statistic in order to test for natural selection by analyzing the several of the highest splits in a coalescent tree. Yang *et al.* (2018) utilizes a reconstructed tree as well as the first few entries of the SFS in order to test whether a given data set was generated under the neutral coalescent. Because these methods analyze topological information, they are also robust to potential confounding by population history. However, our method is different in that it a) allows for both estimation and testing; b) allows for specifying a demographic model, and c) utilizes, we believe, a more accurate probabilistic approximation. We compare these methods with ours in greater detail below.

A third 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 *et al.* 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 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 *et al.* (2019), who proposed 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 *et al.*, 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.

Methods

Our starting point is the standard n-coalescent (Kingman 1982b,a) which is defined as a stochastic process on the set of partitions of the set $\{1, \ldots, n\}$. The process begins at time $t = 0$

in state $\mathcal{C}(0) = \{\{1\}, \ldots, \{n\}\}$. The instantaneous transition rate at time t is $\binom{|\mathcal{C}(t)|}{2}$, where $1 \le |\mathcal{C}(t)| \le n$ 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 $\{\{1, \ldots, n\}\}$. The trajectory of states $\{\mathcal{C}(t) : t \ge 0\}$ can be straightforwardly identified with a bifurcating tree on n leaves, with internal nodes occurring 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. It is shown in Algorithm 1. The main distinction is that the algorithm proceeds *forwards* in time (i.e., from past up to present), as opposed to Kingman's original, retrospective formulation. In these algorithms, $\mathcal{C}_k = \{B_{k,1}, \ldots, B_{k,k}\}$ represents the partition-valued state of the coalescent (resp. splitting) process at level k in the coalescent tree; each $B_{k,i}$ represents a set of nodes that have the same ancestor at level k. It is easy to see that both the forwards- and backwards-in-time algorithms are equivalent. For example, at the top-most level of the tree, m is uniformly distributed on $\{1, \ldots, n-1\}$ (Algorithm 1, line 5), as has been shown for the coalescent (Tajima 1983). More formally, the equivalence follows from e.g. Durrett (2008, Theorem 1.8).

Finally, while Kingman's original model assumes that size of the population is constant, the coalescent can be generalized to allow the effective population size to vary over time by replacing the exponentially-distributed intercoalescence times T_k (line 3 of Algorithm S1) with draws from the waiting time distribution of a point process with rate function $\binom{k}{2}/N_e(t)$ (Griffiths and Tavaré 1994). By defining time on the forwards, as opposed to reverse, axis, the same generalization applies to Algorithm 1. We return to this point in Expected site frequency spectrum.

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

```
1: $\mathcal{C}_1 = \{\{1, \ldots, n\}\}, k = 1$.
2: while $k < n$ do
3:     $T_{k+1} \sim \mathrm{Exp}(k(k+1)/2)$.
4:     Sample $B_{k,i}$ from $\mathcal{C}_k = \{B_{k,1}, \ldots, B_{k,k}\}$ with probability proportional to $(|B_{k,i}| - 1)/(n-k)$.
5:     $m \sim \mathrm{Uniform}(\{1, 2, \ldots, |B_{k,i}| - 1\})$.
6:     Randomly partition $B_{k,i}$ into non-empty subsets $A, A'$ such that $|A| = m$ and $|A'| = |B_{k,i}| - m$.
7:     $\mathcal{C}_{k+1} \leftarrow (\mathcal{C}_k \cup \{A, A'\}) \setminus B_{k,i}$.
8:     $k \leftarrow k + 1$.
   return $T_n, \ldots, T_2, \mathcal{C}_n, \ldots, \mathcal{C}_2$.
```

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 $\{1, \ldots, |B_i| - 1\}$. 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 (A third type of selection, background selection, alters genetic diversity in a way that is indistinguishable from shrinking the effective population size (Charlesworth *et al.* 1993), and is therefore not captured by our approach).

Algorithm 2 β-splitting coalescent model.

```
$\mathcal{C}_1 = \{\{1, \ldots, n\}\}, k = 1$.
while $k < n$ do
    $T_{k+1} \sim \mathrm{Exp}(k(k+1)/2)$.
    Sample $B_{k,i}$ from $\mathcal{C}_k = \{B_{k,1}, \ldots, B_{k,k}\}$ with probability proportional to $(|B_{k,i}| - 1)/(n-k)$.
    $m \sim \mathrm{BetaBinomial}(|B_{k,i}|; \beta, \beta \mid 1 \le m \le |B_{k,i}| - 1)$.
    Randomly partition $B_{k,i}$ into non-empty subsets $A, A'$ such that $|A| = m$ and $|A'| = |B_{k,i}| - m$.
    $\mathcal{C}_{k+1} \leftarrow (\mathcal{C}_k \cup \{A, A'\}) \setminus B_{k,i}$.
    $k \leftarrow k + 1$.
return $T_n, \ldots, T_2, \mathcal{C}_n, \ldots, \mathcal{C}_2$.
```

Such a generalization is shown in Algorithm 2, which is based on a model of cladogenesis proposed by Aldous (1996). Aldous defined a one-parameter family of distributions which he called the *β-splitting model*(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 Discussion). 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

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

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^β_n 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 $\{1, \dots, n-1\}$. For example, Disanto *et al.* (2013) propose a similar class of forward-time Yule models indexed by a discrete parameter ω which bounds the minimal size of any subtree. We prefer the symmetric beta-binomial model because of interpretability, parsimony (it adds only one real-valued—and hence, optimizable—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 to have any evolutionary interpretation (Aldous 1996). We choose to view it empirically as a useful tool for studying natural selection using coalescent-based methods.

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 determine the ESFS under the β-splitting model.

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:

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

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} is the probability that a branch at 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.

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

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

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. Observe that in the preceding equation, the calculation of the expected SFS decomposes into two independent sources of variation: topological variation, captured by $\mathbf{p}^\beta_{nk}$ and depending only on β and combinatorial aspects of the coalescent, and branch length variation, captured by $\mathbb{E}_\eta \mathbf{T}_n$ and depending on the underlying demographic model. In the next subsections, we discuss how to compute these quantities.

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

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

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

$$\mathbb{P}(S_j^k = s_j + 1 \mid \mathbf{S}^{k-1} = \mathbf{s}) = \frac{1}{n-k+1} \times \sum_{\ell=j+1}^{n} (\ell - 1) s_\ell \sum_{q \in \{j, \ell - j\}} f_{k-1,\ell,q}^{\beta}. \quad (9)$$

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

$$\mathbb{P}(S_j^k = s_j - 1 \mid \mathbf{S}^{k-1} = \mathbf{s}) = \frac{j-1}{n-k+1} s_j. \quad (10)$$

In matrix notation, (9) and (10) 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,

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

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 *et al.* (2003) have shown the following relationship for a general size history function η:

$$\mathbb{E}_\eta \mathbf{T}_n = \mathbf{A} \cdot \mathbb{E}_\eta \tilde{\mathbf{T}}_n \quad (11)$$

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

$$\begin{aligned} \mathbb{E}_\eta \tilde{T}_{nj} &= \int_0^\infty \exp\left\{ -\binom{j}{2} R_\eta(t) \right\} \mathrm{d}t, \quad j = 2, \dots, n \quad (12) \\ R_\eta(t) &:= \int_0^t \frac{\mathrm{d}s}{\eta(s)}, \end{aligned}$$

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

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

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 (11) 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 (11) is numerically unstable, so we use a high precision numerical library to evaluate the integral (12) and then (11). 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 genome-wide analysis.

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.

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. 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

$$L(\beta, \theta | \boldsymbol{\phi}) = \|\boldsymbol{\phi}\|_1 \log(\theta \|\mathbb{E}_{\eta,\beta,\cdot}\|_1) - \theta \|\mathbb{E}_{\eta,\beta,\cdot}\|_1 + \frac{\left\langle \boldsymbol{\phi}, \mathbb{E}_{\eta,\beta,\cdot} \right\rangle}{\|\mathbb{E}_{\eta,\beta,\cdot}\|_1}, \quad (13)$$

where the ESFS $\mathbb{E}_{\eta,\beta,\cdot}$ is calculated using the procedure derived in Expected site frequency spectrum. If the mutation rate θ is not known, then the maximum likelihood estimate can be shown to equal

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

Substituting this back into (13), and setting $\mathbf{p} = \boldsymbol{\phi}/\|\boldsymbol{\phi}\|_1$, $\mathbf{q}(\beta) = \mathbb{E}_{\eta,\beta,\cdot}/\|\mathbb{E}_{\eta,\beta,\cdot}\|_1$, we obtain that the profile likelihood

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

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

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 to 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 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

$$\mathbb{P}(\mathcal{T} \mid \beta) = \prod_{k=2}^{n} p_{B_k}^{\beta}(c_k), \tag{14}$$

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

Weighted likelihood When experimenting with this method, we observed a small but consistent performance improvement by reweighting the likelihood (14):

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

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 Topological variance analysis). 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 S9 and S10. The gain was around 0.01–0.04 AUC in each scenario.

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

$$I_c(\mathcal{T}) = \frac{1}{\binom{n-1}{2}} \sum_{t \in \mathcal{T}} |t_r - t_\ell|, \tag{15}$$

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

We also compared our results with the Ω statistic suggested by Li and Wiehe (2013). Specifically, we used $T_3^{(\text{sum})}$, where;

$$T_3^{(\text{sum})} = \sqrt{\frac{12}{4}} \sum_{i=0}^{3} (\widehat{\Omega_i^*} - \frac{1}{2}), \tag{16}$$

(cf. equation (14) in their paper), which the authors found to have good power to detect selection.

Polytomies In practice, we found that current tree inference software often generates multifurcating trees. Since our method assumes a bifurcating tree, we first resolved these polytomies by arbitrarily breaking them into sequences of bifurcation events. Specifically, for each polytomy, we randomly ordered the descendant nodes, and then replaced that polytomy with a sequence of bifurcations in that order. Other methods for breaking polytomies which are biased towards greater or lesser split imbalance are also possible; we verified that our results were not sensitive to this choice (Figures S11 and S12). Of course, polytomies could well represent additional signals of selection. Our current implementation ignores this, but we discuss potential extensions in Discussion.

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 The β-splitting family, 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:

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

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

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

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.

Data analysis pipeline

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

Results

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

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

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

where $\mathbf{p}^\beta_{nk}$ was defined in Dynamic programming algorithm for $\mathbf{p}^\beta_{nk}$, and the second inequality holds because $\mathbb{E}d_k = n/k$ under any leaf-exchangeable tree distribution. Computing $\mathrm{var}_\beta(d_k)$ 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

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

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

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

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

$$\mathrm{var}_\beta(d_2) = \mathrm{var}(X \mid 1 \le X \le n-1)$$

where $X \sim \mathrm{BetaBinomial}(n; \beta, \beta)$.

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

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

We further define

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

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 $f_\Omega(\theta l)$.

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 S1). 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 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 distributions at $\beta = 0$ are closer to the $\beta \to \infty$ limit than the $\beta \to -2$ limit.

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 (13) over β with respect to the observed frequency spectrum. btree is the tree-sequence based estimate, obtained by maximizing the conditional likelihood defined in (14) over β conditional on a given tree. As a baseline, we also compared our method to Colless' statistic, Li and Wiehe's Omega (see From inferred trees) 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 Simulation studies.

Directional selection We simulated a single population with constant population size $N = 2 \times 10^4$. The simulated region was $L = 10^5$ base pairs (bp), with recombination and mutation rates of 1.25×10^{-8} and 2.5×10^{-8} per bp 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 before present into the middle of the region, and we restarted the simulation if the mutation was lost or fixed. Following Stern *et al.* (2019), we varied two parameters: the selection coefficient $s \in \{.001, .003, .01, .02\}$, and the allele frequency $F \in \{0.25, 0.5, 0.75\}$ of the mutation when the simulation terminated. Genic selection was assumed, i.e. the relative fitnesses of the wild-type homozygotes, heterozygotes, and derived homozygotes were 1, $1 + s/2$, and $1 + s$, respectively.

Figure 1 displays ROC curves and AUC scores when $s = 0.01$. (Results for all scenarios are shown in Figure S6.) 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 the frequency spectrum (Kaplan *et al.* 1989). Among the frequency spectrum-based methods, ours (btree) achieves the best AUC

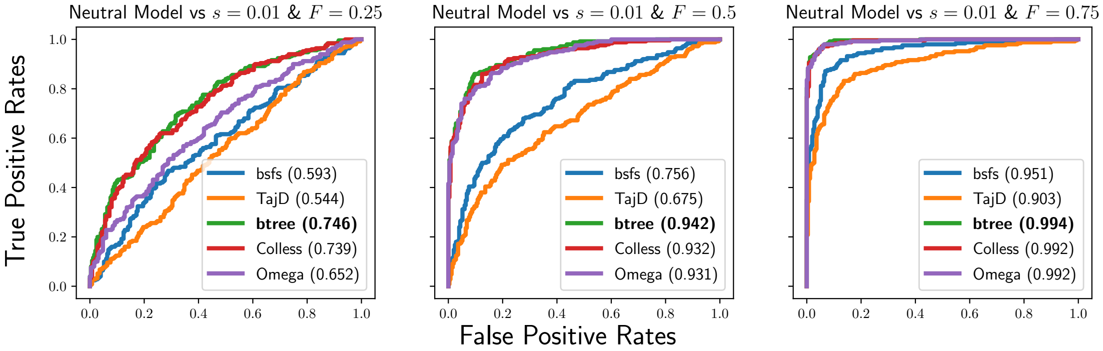

Figure 1 ROC curves for positive genic selection. s represents the selection coefficient and F represents allele frequency of the mutation in the sample. AUC is shown in parentheses. bsfs and btree are our SFS- and tree-based methods, respectively. TajD is Tajima's D. Colless' and Omega are defined in Related tree imbalance statistics.

across all settings of F. Omega has low AUC when the beneficial mutation has low sample frequency ($F = 0.25$). bsfs outperforms Tajima's D in all settings. The results were consistent across other selection coefficients (Figure S6). btree achieves at least 0.8 AUC for $s \geq 0.003$ and $F \geq 0.5$. It has significantly higher AUC than Colless' and Omega (vs Colless': $p = .0048$; vs Omega: $p = .023$, Wilcoxon signed rank test), but the overall gain is small (vs Colless': mean ΔAUC $= 0.0051$, vs Omega: mean ΔAUC $= 0.023$). 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 ΔAUC $= 0.049$).

Finally, in Table S3 we computed the type I error at the nominal level $\alpha = 0.05$ for each method on the set of neutral simulations, with p-values computed using the normal approximation described below. If we knew the exact distribution of each test statistic under the null, then all of these entries would be approximately $\alpha = 0.05$, so this table measures the adequacy of our approximation. The table indicates that the p-values are fairly well calibrated, with bsfs and Omega having slightly inflated type I error, while the other methods are slightly conservative.

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. Starting at 4000 generations before the end of the simulation, heterozygous advantageous mutations were introduced at a constant rate. 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 2 shows ROC and AUC scores for each of the methods. In contrast to the case of directional selection, SFS-based statistics did better than tree-based statistics in this example. bsfs has higher AUC values for all three simulations. All tree-based statistics perform similarly. The results were consistent if we varied the starting generation when mutations were introduced (Figure S8). btree outperforms Colless' statistic, but not the Omega, however, the differences are slight (vs. Colless': mean ΔAUC $= 0.0028$, vs. Omega: mean ΔAUC $= -0.0076$) and not significant (vs Colless': $p = .38$; vs Omega: $p = .91$, Wilcoxon signed rank test). Among the SFS-based statistics, bsfs achieved significantly higher AUC ($p = 0.0011$, Wilcoxon signed rank test) than Tajima's D with a mean ΔAUC $= 0.022$. 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.

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 (Expected site frequency spectrum).

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

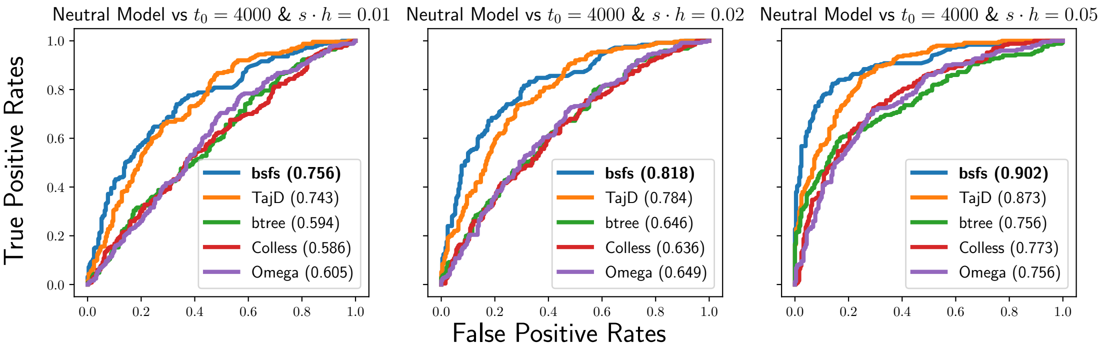

Figure 2 ROC curves for advantageous heterozygote mutation simulations. s is the selection coefficient, h is the dominance factor, and t_0 is the number of generations before present when the mutation was introduced into the sample. Other abbreviations are the same as in Figure 1.

rate of exponential growth, s is the selective coefficient of the beneficial mutation and h is the dominance parameter.

In Figure 3a, 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 (rightmost 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, 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 3b, 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 by simulating four 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 4b, 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

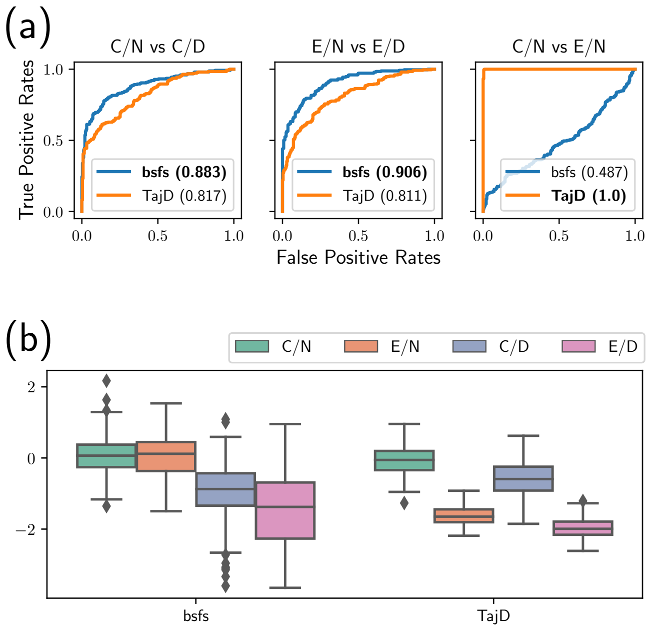

Figure 3 Classifying directional selection under different growth scenarios: **C**onstant vs. **E**xponential and **N**eutral vs. **D**irectional. (a) bsfs is better at detecting true signals in the first two figures. In the third panel, D conflates exponential growth with selection. (b) Under neutrality, bsfs has a zero-centered empirical distribution regardless of growth scenario, whereas the distribution of D is shifted.

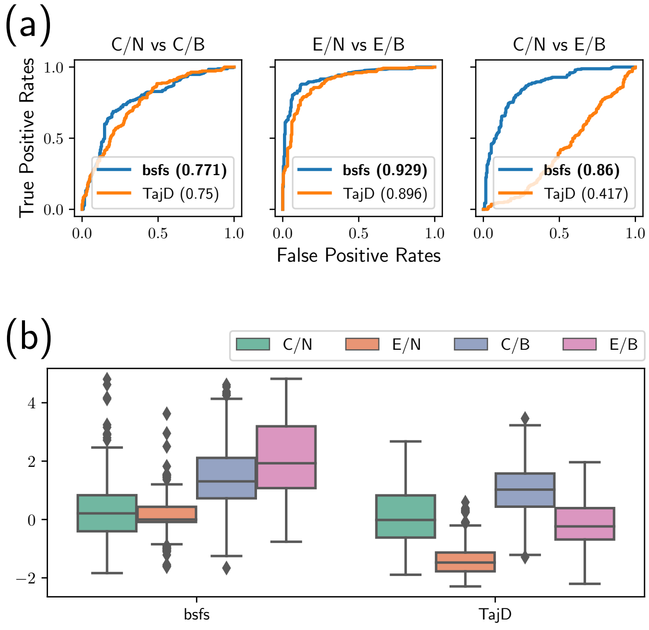

Figure 4 Classifying balancing selection under different growth scenarios. Abbreviations follow the same convention as in Figure 3. (a) bsfs performs better for detecting true signals in the first two figures. In the third figure Tajima's D fails to detect selection. (b) Under neutrality, bsfs has a zero-centered empirical distribution and balancing selection shifts the distribution upward. Balancing selection shifts D to positive values, but exponential growth pulls it downward.

neutrality with constant size. In contrast, the null distribution of bsfs is invariant after correcting for demography.

Testing procedure Finally, to give intuition for how the testing procedure used in Real data analysis works, we simulated a longer chromosome ($L = 2.5 \times 10^7$ bp) with a beneficial mutation introduced in the middle position. The population size, the mutation rate and the recombination rate were the same as in the previous simulations. We only focused on the case where the selection coefficient $s = 0.01$ and sample allele frequency of the mutation was $F = 0.75$. As detailed in Data analysis pipeline, we applied a change point detection algorithm to create segments of estimates along the genome (Combining $\hat{\beta}$ from multiple segments). This has the effect of smoothing nearby similar estimates, thereby reducing noise and false discoveries. We then computed approximate p-values using a Gaussian approximation (Significance testing).

Figure 5 (excerpted from Figure S7) shows the segments resulting from this procedure over 250 simulations for bsfs and Tajima's D. Both methods are enriched for small p-values near the site of the mutation (highlighted in gray), but moreso with bsfs. After performing the segmentation, we classified all those segments which overlapped the selected locus as "under selection", and all other segments were classified as "neutral". Neutral segments below the rejection threshold therefore represent false discoveries. In Figure S3, we varied the p-values for each method to see how many true discoveries result as a function of the number of false discoveries. Tree-based methods yield consistently better results overall.

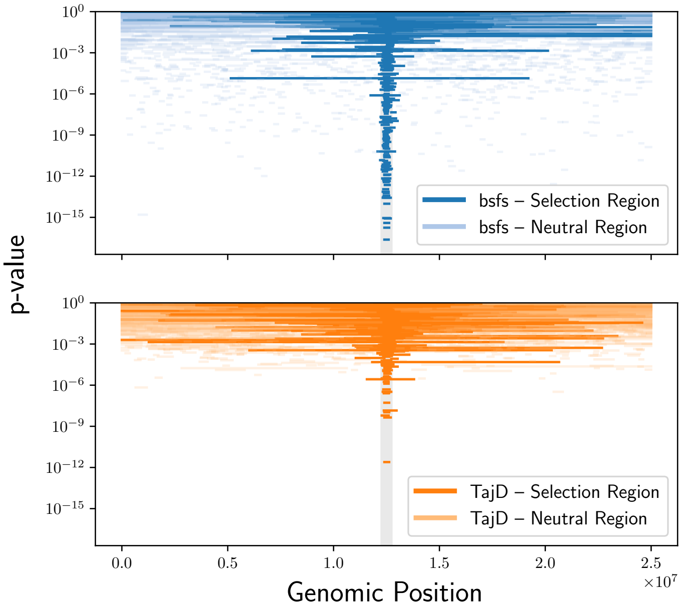

Figure 5 Scan statistics for 250 simulations with a single beneficial mutation at the middle position (highlighted in gray and enlarged for clarity) with selection coefficient $s = 0.01$ and sample allele frequency $F = 0.75$. Both bsfs and Tajima's D have smaller p-values around the mutation, but the signal in bsfs is more pronounced. Neutral segments are shown with transparency to reduce overplotting.

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, *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. 20).

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 (The β-splitting family). 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 in 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.

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 S13 and S14).

Balancing selection Next, we used our method to study long-term balancing selection in 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.

Results of genome-wide scan In Supplemental Tables, we list the genome-wide 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 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.

Discussion

In this paper, we presented some new methods to detect natural selection by generalizing Kingman's coalescent to allow for 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 S6 and S8 with Figures 3 and 4 of Stern *et al.* (2019). However, an advantage of our method is that it 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 S2). 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 from our method can 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 searching through genome-wide scans performed using our β estimates, along with classical neutrality tests, across all populations.

There are several ways our model could be improved. We focused on 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. A drawback of our method is that the current implementation cannot model ancestral population structure. In particular, barriers to gene flow between ancestral populations could be confused with balancing selection. Using techniques developed by Kamm *et al.* (2017, 2020), our method could potentially be extended to model this, though such an extension would be non-trivial. Another, related, 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 *et al.* 2014; Terhorst *et al.* 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 From inferred trees. We experimented with this, but found that the branch lengths from 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.

On a similar note, when running our method on tree sequence data, we observed that the 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, polytomies could comprise 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 *et al.* 2016; Blath *et al.* 2016). In the future, our method could be extended to work under this more general model.

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.

Acknowledgments

This project was conceived during the workshop *Ecosystem dynamics: stakes, data and models* at Université Paris–Saclay. JT thanks the Institut Pascal for organizing the event, travel funding, and the limitless supply of M&M's; and Amandine Véber and Raazesh Sainudiin for helpful early discussions.

Funding

This research was supported by NSF grant DMS-2052653.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. Genetics. 183:249–258.

Aldous D. 1996. Probability Distributions on Cladograms. In: Aldous D, Pemantle R, editors, *Random Discrete Structures*. The IMA Volumes in Mathematics and its Applications. pp. 1–18. New York, NY. Springer.

Barrett RD, Schluter D. 2008. Adaptation from standing genetic variation. Trends in ecology & evolution. 23:38–44.

Berg JJ, Coop G. 2015. A coalescent model for a sweep of a unique standing variant. Genetics. 201:707–725.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. The American Journal of Human Genetics. 74:1111–1120.

Bhaskar A, Wang YXR, Song YS. 2015. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Research. 25:268–279.

Biswas S, Akey JM. 2006. Genomic insights into positive selection. TRENDS in Genetics. 22:437–446.

Blath J, Cronjäger MC, Eldon B, Hammer M. 2016. The site-frequency spectrum associated with Ξ-coalescents. Theoretical Population Biology. 110:36–50.

Blum MG, François O. 2006. Which random processes describe the tree of life? a large-scale study of phylogenetic tree imbalance. Systematic Biology. 55:685–691.

Botchkarev VA, Fessing MY. 2005. Edar Signaling in the Control of Hair Follicle Development. Journal of Investigative Dermatology Symposium Proceedings. 10:247–251. Publisher: Elsevier.

Branicki W, Brudnik U, Draus-Barini J, Kupiec T, Wojas-Pelc A. 2008. Association of the SLC45A2 gene with physiological human hair colour variation. Journal of Human Genetics. 53:966–971. Number: 11 Publisher: Nature Publishing Group.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS genetics. 5:e1000336.

Celisse A, Marot G, Pierre-Jean M, Rigaill GJ. 2018. New efficient algorithms for multiple change-point detection with reproducing kernels. Computational Statistics & Data Analysis. 128:200–220.

Charlesworth B, Morgan M, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134:1289–1303.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2:e64.

Disanto F, Schlizio A, Wiehe T. 2013. Yule-generated trees constrained by node imbalance. Mathematical Biosciences. 246:139–147.

Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Barta C, Lu RB, Zhukova OV, Kim JJ, Siniscalco M, New M *et al.* 2012. A global view of the OCA2-HERC2 region and pigmentation. Human Genetics. 131:683–696.

Durrett R. 2008. *Probability Models for DNA Sequence Evolution*. Springer, New York. second edition.

Fay JC, Wu CI. 2000. Hitchhiking under positive darwinian selection. Genetics. 155:1405–1413.

Feder AF, Kryazhimskiy S, Plotkin JB. 2014. Identifying signatures of selection in genetic time series. Genetics. 196:509–522.

Ferretti L, Ledda A, Wiehe T, Achaz G, Ramos-Onsins SE. 2017. Decomposing the Site Frequency Spectrum: The Impact of Tree Topology on Neutrality Tests. Genetics. 207:229–240.

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI *et al.* 2016. Detection of human adaptation during the past 2000 years. Science. 354:760–764. Publisher: American Association for the Advancement of Science Section: Report.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations.

Genetics. 133:693–709.

Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, Thomas MG. 2011. Evolution of lactase persistence: an example of human niche construction. Philosophical Transactions of the Royal Society B: Biological Sciences. 366:863–877.

Griffiths R, Tavaré S. 1998. The age of a mutation in a general coalescent tree. Communications in Statistics. Stochastic Models. 14:273–295.

Griffiths RC, Tavaré S. 1994. Sampling theory for neutral alleles in a varying environment. Proc. R. Soc. London B.. 344:403–410.

Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. 2019. Tree-sequence recording in slim opens new horizons for forward-time simulation of whole genomes. Molecular ecology resources. 19:552–566.

Haller BC, Messer PW. 2019. SLiM 3: Forward Genetic Simulations Beyond the WrightFisher Model. Molecular Biology and Evolution. 36:632–637.

Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC, Kidd JR, Kidd KK. 2007. Evidence of Positive Selection on a Class I ADH Locus. American Journal of Human Genetics. 80:441–456.

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular ecology. 22:4606–4618.

Hlusko LJ, Carlson JP, Chaplin G, Elias SA, Hoffecker JF, Huffman M, Jablonski NG, Monson TA, ORourke DH, Pilloud MA *et al*. 2018. Environmental selection during the last ice age on the mother-to-infant transmission of vitamin D and fatty acids through breast milk. Proceedings of the National Academy of Sciences. 115:E4426–E4432. Publisher: National Academy of Sciences Section: PNAS Plus.

Kamm J, Terhorst J, Durbin R, Song YS. 2020. Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. Journal of the American Statistical Association. 115:1472–1487.

Kamm JA, Terhorst J, Song YS. 2017. Efficient computation of the joint sample frequency spectra for multiple populations. Journal of Computational and Graphical Statistics. 26:182–194.

Kaplan NL, Darden T, Hudson RR. 1988. The coalescent process in models with selection. Genetics. 120:819–829.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. Genetics. 123:887–899.

Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. 2019a. Inferring whole-genome histories in large population datasets. Nature Genetics. 51:1330–1338.

Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. 2019b. Inferring whole-genome histories in large population datasets: inferred tree sequences for 1000 Genomes. https://zenodo.org/record/3051855.

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics. 167:1513–1524.

Kingman JFC. 1982a. The coalescent. Stoch. Process. Appl.. 13:235–248.

Kingman JFC. 1982b. On the genealogy of large populations. J. Appl. Prob.. 19A:27–43.

Krone SM, Neuhauser C. 1997. Ancestral processes with selection. Theoretical Population Biology. 51:210–237.

Lehmann EL, Casella G. 2006. *Theory of point estimation*. Springer Science & Business Media.

Li H. 2011. A New Test for Detecting Recent Positive Selection that is Free from the Confounding Impacts of Demography. Molecular Biology and Evolution. 28:365–375.

Li H, Wiehe T. 2013. Coalescent Tree Imbalance and a Simple Test for Selective Sweeps Based on Microsatellite Variation. PLOS Computational Biology. 9:e1003060. Publisher: Public Library of Science.

Liu X, Ong RTH, Pillai EN, Elzein AM, Small KS, Clark TG, Kwiatkowski DP, Teo YY. 2013. Detecting and Characterizing Genomic Signatures of Positive Selection in Global Populations. The American Journal of Human Genetics. 92:866–881.

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N *et al*. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 7:e1002326.

Malaspinas AS, Malaspinas O, Evans SN, Slatkin M. 2012. Estimating allele age and selection coefficient from time-serial data. Genetics. 192:599–607.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet. Res., Camb.. 23:23–35.

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5:e1000471.

Mooers A, Heard S. 1997. Inferring Evolutionary Process from Phylogenetic Tree Shape. Quarterly Review of Biology. 72:31–54.

Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. Genetics. 145:519–34.

Neyman J, Pearson ES. 1933. Ix. on the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character. 231:289–337.

Palamara PF, Terhorst J, Song YS, Price AL. 2018. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. Nature Genetics. 50:1311–1317.

Pitman J. 1999. Coalescents with multiple collisions. Annals of Probability. 27:1870–1902.

Polanski A, Bobrowski A, Kimmel M. 2003. A note on distributions of times to coalescence, under time-dependent population size. Theoretical Population Biology. 63:33–40.

Polanski A, Kimmel M. 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics. 165:427–436.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genetics. 10:e1004342.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science. 312:1614–20.

Sagitov S. 1999. The general coalescent with asynchronous mergers of ancestral lines. Journal of Applied Probability. 36:1116–1125.

Sainudiin R, Véber A. 2016. A beta-splitting model for evolutionary trees. Royal Society open science. 3:160016.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics. 132:1161–1176.

Schweinsberg J. 2003. Coalescent processes obtained from supercritical galton–watson processes. Stochastic processes and their Applications. 106:107–139.

Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics. 51:1321–1329.

Spence JP, Kamm JA, Song YS. 2016. The site frequency spectrum for general coalescents. Genetics. 202:1549–1561.

Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Molecular Biology and Evolution. 22:63–73.

Stern AJ, Nielsen R. 2019. Detecting natural selection. Handbook of Statistical Genomics: Two Volume Set. pp. 397–40.

Stern AJ, Wilton PR, Nielsen R. 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLOS Genetics. 15:e1008384. Publisher: Public Library of Science.

Tajima F. 1983. Evolutionary relationship of dna sequences in finite populations. Genetics. 105:437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics. 123:585–595.

Takahata N. 1993. Allelic genealogy and human evolution. Mol. Biol. Evol.. 10:2–22.

Terhorst J, Schlötterer C, Song YS. 2015. Multi-locus analysis of genomic time series data from experimental evolution. PLoS Genet. 11:e1005069.

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

Truong C, Oudre L, Vayatis N. 2020. Selective review of offline change point detection methods. Signal Processing. 167:107299.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annual review of genetics. 47:97–120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biology. 4:e72.

Yang Z, Li J, Wiehe T, Li H. 2018. Detecting Recent Positive Selection with a Single Locus Test Bipartitioning the Coalescent Tree. Genetics. 208:791–805.

Yasumizu Y, Sakaue S, Konuma T, Suzuki K, Matsuda K, Murakami Y, Kubo M, Palamara PF, Kamatani Y, Okada Y. 2020. Genome-Wide Natural Selection Signatures Are Linked to Genetic Risk of Modern Phenotypes in the Japanese Population. Molecular Biology and Evolution. 37:1306–1316.

Zhao B, Luo T, Li T, Li Y, Zhang J, Shan Y, Wang X, Yang L, Zhou F, Zhu Z *et al*. 2019. 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. Nature Genetics. 51:1637–1644. Number: 11 Publisher: Nature Publishing Group.

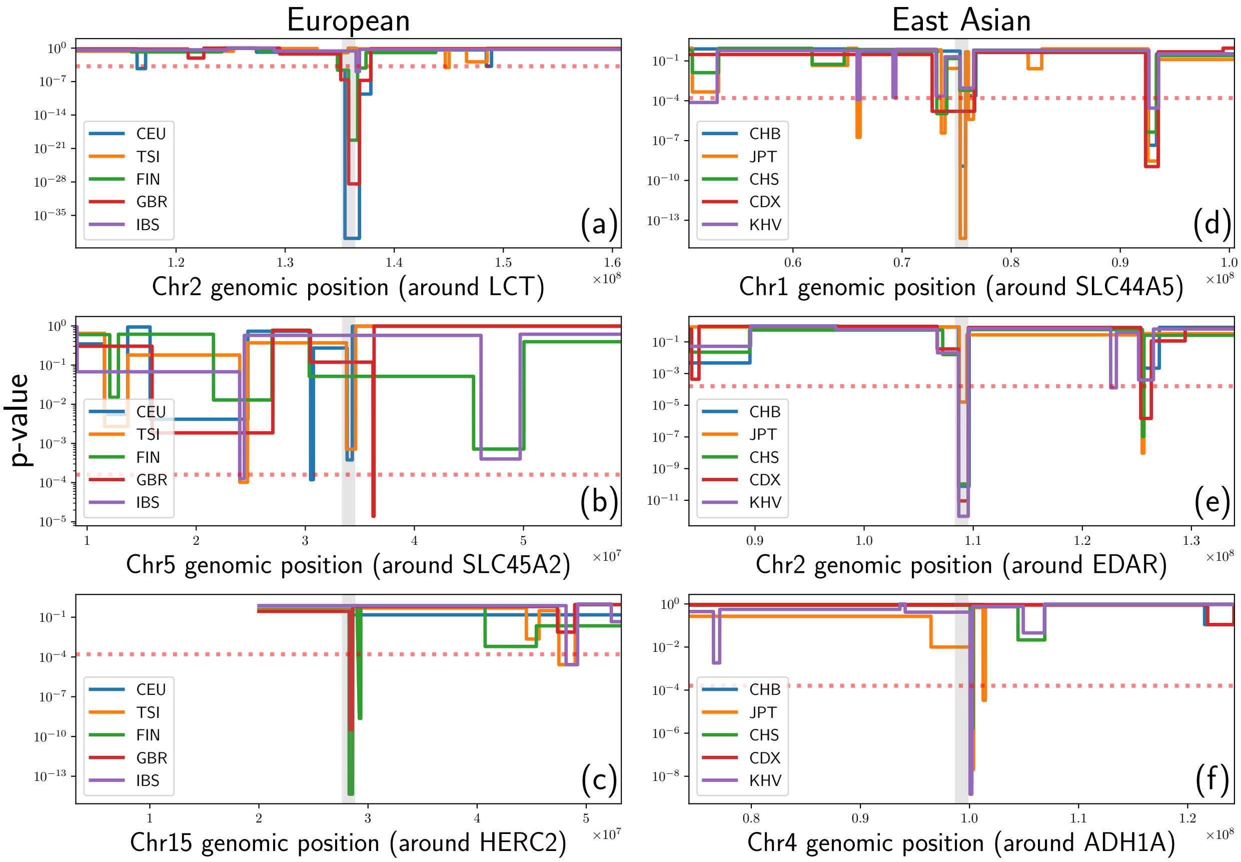

Figure 6 Results of directional selection p-value scan for 1000 Genomes Project using median centered btree (Median-centered estimates of β). The Bonferonni-corrected significance level is 1.6×10^{-4} (Red dashed line). Significant populations for each gene: (a) CEU, GBR, FIN; (b) None; (c) FIN, CEU, 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.

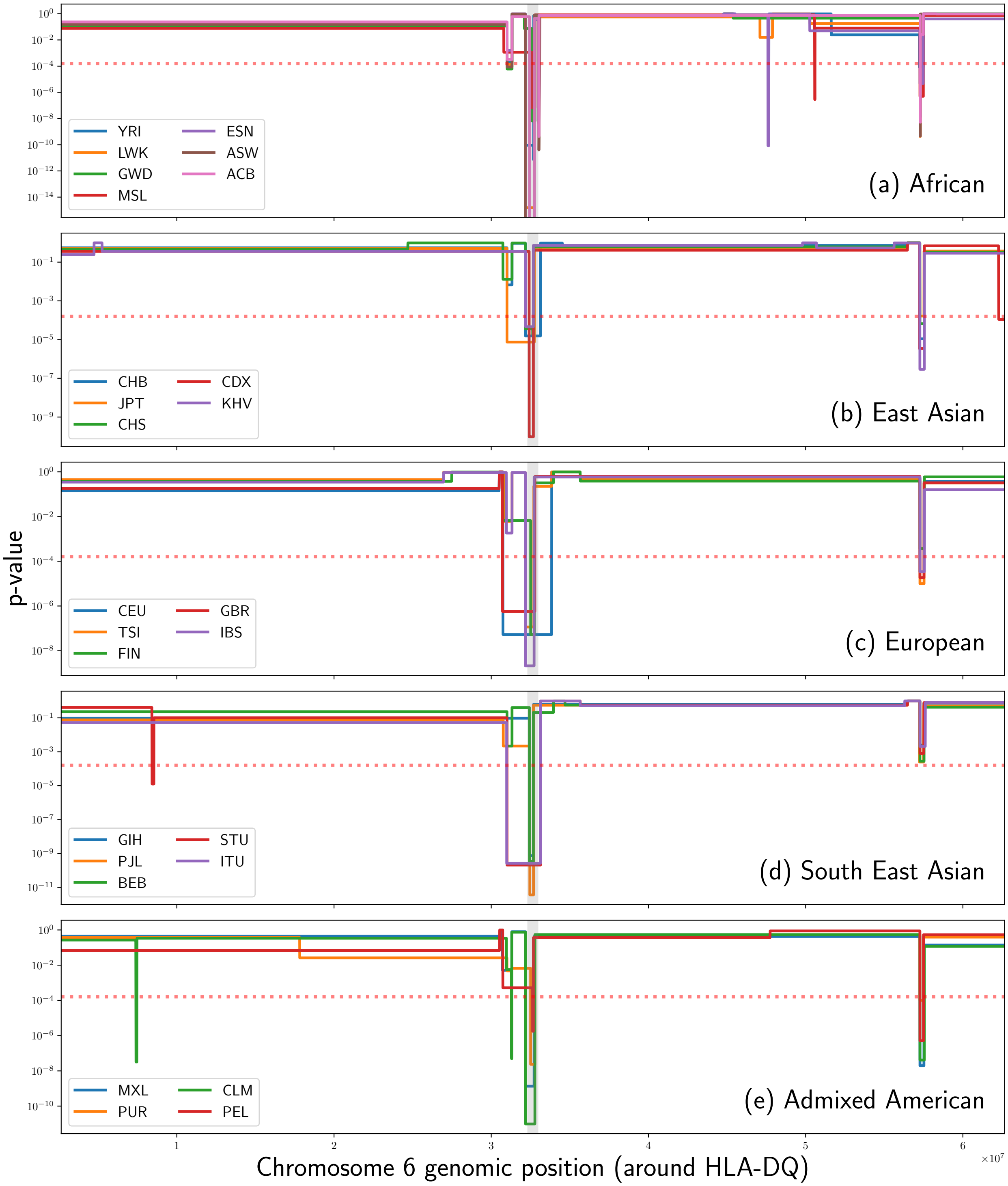

Figure 7 Genome Scan p-values of the bsfs segments around *HLA-DQ*. Most of the 1000 Genomes subpopulations have a pronounced balancing selection signal in this region.