Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests

Luca Ferretti[1,*], Alice Ledda[2], Thomas Wiehe[3], Guillaume Achaz[4,5,6] and Sebastian E. Ramos-Onsins[7]

(1) The Pirbright Institute, Woking, UK. (2) Department of Infectious Disease Epidemiology, Imperial College, London. (3) Institute of Genetics, University of Cologne. (4) Evolution Paris-Seine (UMR CNRS 7138), UPMC, Paris. (5) Atelier de Bio-Informatique, UPMC, Paris. (6) Stochastic Models for the Inference of Life Evolution, CIRB (UMR INSERM 7241), Collège de France, Paris. (7) CRAG, Bellaterra, Barcelona, Spain.

Abstract

We investigate the dependence of the site frequency spectrum (SFS) on the topological structure of genealogical trees. We show that basic population genetic statistics – for instance estimators of θ or neutrality tests such as Tajima's D – can be decomposed into components of waiting times between coalescent events and of tree topology. Our results clarify the relative impact of the two components on these statistics. We provide a rigorous interpretation of positive or negative values of an important class of neutrality tests in terms of the underlying tree shape. In particular, we show that values of Tajima's D and Fay and Wu's H depend in a direct way on a peculiar measure of tree balance which is mostly determined by the root

*Email: luca.ferretti@gmail.com

balance of the tree. We present a new test for selection in the same class as Fay and Wu's H and discuss its interpretation and power. Finally, we determine the trees corresponding to extreme expected values of these neutrality tests and present formulae for these extreme values as a function of sample size and number of segregating sites.

Introduction

Coalescent theory (Kingman, 1982; Hein *et al.*, 2004; Wakeley, 2009) provides a powerful framework to interpret the mutation patterns in a sample of DNA sequences. Grounded in the neutral theory of molecular evolution (Kimura, 1985), binary coalescent trees are the dual backward representations of the continuous-forward-time diffusion model of genetic drift. In this view, sequences are related by a genealogical tree where leaf nodes represent the sampled sequences at present time and internal nodes (coalescent events) represent last common ancestors of the leaves underneath. In particular, the root node represents the most recent common ancestor of the whole sample.

In species phylogeny and epidemiology, tree structure is often used to compare different models of evolution or to fit model parameters (Bouckaert *et al.*, 2014). Two summary statistics are routinely used to characterize tree structure: the γ statistic relates to the waiting times (Pybus *et al.*, 2000) and the β statistic to tree balance (Blum and François, 2006). Importantly, these statistics can only be computed after the tree structure was independently inferred – typically by phylogenetic reconstruction methods (Felsenstein, 2004).

In population genetics, the historical relationship among non-recombining sequences is represented by a single genealogical tree. The tree is completely determined by the waiting times and the branching order of coalescent events. The waiting times determine branch lengths, the branching order determines tree shape. Population genetic statistics, such as estimates of the scaled mutation rate or tests of the neutral evolution hypothesis (neutrality tests) are sensitive to waiting times and tree shape.

The site frequency spectrum (SFS) is one of the most used statistics in population genetics. The unfolded site frequency spectrum $\boldsymbol{\xi} = (\xi_1, ..., \xi_{n-1})$ of a sample of n sequences is defined as the vector of counts ξ_i, $i \in \{1, ..., n-1\}$, of all polymorphic sites with a derived allele ("mutation") at frequency i/n. The SFS is a function of both tree structure and mutational process. For a given mutational process, the SFS carries information on the underlying, but not directly observable, genealogical trees and therefore on the forward process that has generated the trees. For a non-recombining locus, the SFS carries information on the realized coalescent tree and can be used to estimate tree structure (both waiting times and topology).

Variation over time in the effective population size affects the expected waiting times between coalescent events. In the past much attention in theoretical works has been paid to the relation between waiting times and population size variation. For example, skyline plots (Pybus *et al.*, 2000) are directly used to infer variation of population size (Ho and Shapiro, 2011) although care should be taken while using this approach (Lapierre *et al.*, 2016). More generally, formulae of the SFS can be generalized to include deterministic changes of population size (Griffiths and Tavaré, 1998; Zivkovic and Wiehe, 2008; Liu and Fu, 2015). In contrast, the influence of tree shape on the SFS has not yet been tackled analytically.

The shape of a tree can range from completely symmetric trees, in which all internal nodes evenly split the lineages, to caterpillar trees, in which each node isolates exactly one lineage. In the standard neutral model – as well as in any other equal-rate-Markov (ERM) or Yule model (Yule, 1925) – both of these extreme cases are very unlikely to appear by chance (Blum and François, 2006). In fact, since the number of binary tree shapes (enumerated by the Wedderburn-Etherington numbers, Sloane and Plouffe (1995)) grows rapidly with the number of sequences n, any specific tree shape is arbitrarily improbable if n is sufficiently large. Nonetheless, tree topology is a major determinant of the SFS. For example, a caterpillar shape leads to a large excess of singleton mutations, while a completely symmetric tree leads to an over-representation of intermediate frequency alleles.

This study aims at a providing a systematic analysis of the impact of the structure of genealogical trees upon the SFS. First, we introduce the theoretical framework for neutrality tests and tree balance. In particular, we develop a new measure of imbalance appropriate for population genetics. Then, we present the decomposition of the SFS in terms of waiting times and tree shape. We discuss the case of a single non-recombining locus, assuming a single realized tree (fixed topology). As recombination affects mostly lower branches of the tree, this constitutes also an excellent approximation for a locus with a low level of recombination.

We present a mathematically rigorous, yet intuitive interpretation of neutrality tests in terms of tree topology and branch lengths. We focus on a subclass of tests of special interest and simplicity. A qualitative summary of the results about the interpretation of neutrality tests is given in Table 4. We also propose a new neutrality test L for selection. Finally, we find the trees corresponding to the maximum and minimum expected values of the tests and provide explicit formulae for these extreme values.

Methods

Trees can be divided into time segments ("levels") delimited by the nodes. Each level is unambiguously characterized by its number of lineages k, $2 \leq k \leq n$. The most recent level has n lineages, the most ancient level (from the root to the next internal node) has 2 lineages. Hereafter, the branches and internal nodes close to the root will be referred to as 'upper part' of the tree; conversely, the 'lower part' is close to the leaves.

The waiting times between subsequent "binary" coalescent events, i.e. the level heights, are denoted by t_k. For trees with coalescent events involving multiple mergers, some of the "binary" waiting times could be null, i.e. $t_k = 0$. For example, if four lineages would coalesce together in a tree with five lineages, and then the two remaining lineages would coalesce to form the root, then $t_3 = 0$.

In a neutral, panmictic population of ploidy p (typically $p = 1$ or 2) and constant effective population size N_e that can be modelled by the Kingman coalescent,

the t_k are exponentially distributed with parameter $k(k-1)$, when the time is measured in $2pN_e$ generations (Wakeley, 2009). Two summary tree statistics are the height $h = \sum_{k=2}^{n} t_k$, that is the time from the present to the most recent common ancestor, and the total tree length $l = \sum_{k=2}^{n} kt_k$. Basic coalescent theory states $E(h) = 1 - 1/n$ and $E(l) = a_n$, where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ is the $(n-1)$th harmonic number.

Tree imbalance per level

Following Fu (1995), we define the *size* d_k of a branch from level k as the number of leaves that descend from that branch. Any mutation on this branch is carried by d_k sequences from the present sample. We denote by $P(d_k = i|T)$ the probability that a randomly chosen branch of level k is of size i, given tree T. The complete set of distributions $P(d_k = i|T)$ for each i and k determines uniquely the shape of the tree T.

The mean number of descendants across all branches from level k is $E(d_k) = \sum_{i=1}^{n-k+1} iP(d_k = i|T) = n/k$. This holds for any tree, since all n present-day sequences must descend from one of the k branches from that level.

In contrast, the size variance, $\mathrm{Var}(d_k)$, depends on the tree topology: at all levels, it is almost zero in completely balanced trees and maximal in caterpillar trees, where all nodes isolate one leaf from the remaining subtree. For this reason, we propose the variance $\mathrm{Var}(d_k)$ as the natural measure of imbalance for each level.

The bounds on $\mathrm{Var}(d_k)$ - shown in Figure 1 - vary greatly from level to level: for example, the variance of the uppermost level $\mathrm{Var}(d_2) \in [0, (n/2-1)^2]$, whereas $\mathrm{Var}(d_n) = 0$ (since $d_n = 1$ for all branches). More generally, the maximum variance at a given level k is obtained in trees where $k-1$ lineages lead to exactly one leaf and one lineage has $n-k+1$ descendants. For this case, we compute

$$\max_T \mathrm{Var}(d_k) = \frac{k-1}{k}\, 1^2 + \frac{1}{k}\,(n-k+1)^2 - \left(\frac{n}{k}\right)^2 = (k-1)\left(\frac{n}{k} - 1\right)^2 . \quad (1)$$

Minimum variance at level k is obtained when all lineages have either[1] $\lfloor n/k \rfloor$ or

[1]We denote by $\lfloor x \rfloor$ the floor of x, i.e the largest integer smaller or equal to x.

$\lfloor n/k \rfloor + 1$ descendants and it is always $\leq 1/4$:

$$\min_T \mathrm{Var}(d_k) = (n/k - \lfloor n/k \rfloor)\ (\lfloor n/k \rfloor + 1 - n/k) \leq \frac{1}{4}\,. \tag{2}$$

Tree imbalance in population genetics versus phylogenetics

Measures of tree topology, and especially tree imbalance, have received considerable attention in the phylogenetic literature (Blum and François, 2006). Several measures of imbalance have been proposed, among them the Sackin's and Colless' indices (Blum and François, 2005; Blum *et al.*, 2006), which depend only on tree topology and not on branch lengths. In the context of phylogenies of genes from different species, the divergence is expected to be large enough such that there would be substitutions on all branches to resolve - in principle - the tree topology. Some of these substitutions are expected to have a functional role (e.g. non-synonymous substitutions, indels). Therefore, almost all splits between lineages should be detectable and correspond to a functional/phenotypic difference between species.

In theory, the same statistics could be applied to the genealogy of a single population or a sample. However, genealogical trees in population genetics are usually much shorter than phylogenetic trees. For short non-recombining sequences, there could be many branches without any mutation event on them. This raises two issues: the detection of imbalance in trees with short branches, and its evolutionary meaning.

Regarding detection, neither a mutation-free branch nor the split above it could be detected from sequences. Hence, a split should be weighted by the probability of being detected through sequence comparison. For example, let I_k be a measure of imbalance at the kth level. Using the probability that there is at least one mutation on level k, that is $1 - e^{-\theta k t_k}$, as a weight function, the combined statistics becomes

$$\bar{I} = \frac{\sum_{k=2}^{n}(1 - e^{-\theta k t_k}) I_k}{\sum_{k=2}^{n}(1 - e^{-\theta k t_k})} \approx \frac{\sum_{k=2}^{n} \theta k t_k I_k}{\sum_{k=2}^{n} \theta k t_k} = \frac{1}{l}\sum_{k=2}^{n} k t_k I_k \quad \text{for small } \theta. \tag{3}$$

On the other hand, from an evolutionary point of view, the importance of a given branch - and of the adjacent splits in the tree - is related to the number of

mutations on the branch. For example, consider a branch that is not supported by any mutation. Its significance for future evolution is null, since there is no selective difference between identical alleles and there is no effect on the genetic variability of the population. This branch could be contracted to zero length, and the splits collapsed into a polytomy of three lineages, without any effect on the present population or on future evolution. (Even a branch that is supported only by non-epistatic neutral mutations does not affect in any way the future selective processes, even if it has an impact on the genetic diversity of the population.) Since mutation-free branches do not have evolutionary significance, their weight in imbalance measures should be low.

Since selective effects and effects on genetic diversity are both proportional to the number of mutations along the branches, it seems reasonable to weight local imbalance measures by the expected number of mutations in the branches supporting them. For example, a measure of imbalance I_k for each level would be weighted by the expected number of mutations $\theta k t_k$ at that level. In this case, we obtain the same statistics $\bar{I} = \sum_{k=2}^{n} k t_k I_k / l$ as in equation (3) above.

A new measure of tree imbalance

We propose an informative statistics on tree balance based on $\mathrm{Var}(d_k)$ and the reasoning in the previous section. We can compute the variance in branch size for each level of the tree, then average it across levels. Fixing a tree T, the average variance in branch size across all levels k is

$$\overline{\mathrm{Var}(d)} = \frac{\sum_{k=2}^{n} k\, t_k \mathrm{Var}(d_k)}{\sum_{k=2}^{n} k\, t_k} = \frac{1}{l} \sum_{k=2}^{n} k\, t_k \mathrm{Var}(d_k)\,. \tag{4}$$

This summary statistic contains the natural weights $k\, t_k / l$, that is the fraction of branch lengths at level k, discussed in the previous section. Note that this average is different from the total variance in offspring number, *i.e.* when the variance of sizes is taken across all branches, irrespective of their level. The statistics $\overline{\mathrm{Var}(d)}$ corresponds instead to the "within-level" component of variance.

To better understand $\overline{\mathrm{Var}(d)}$, we study the extent to which each level contributes to the statistics. Figure 1 shows contributions per unit time and per

whole level. In the first case, $\mathrm{Var}(d_k)$ are weighted by the number of lineages k, while in the second they are weighted by the length at level k, $k\mathrm{E}(t_k)$, which is $1/(k-1)$ for constant population size. Figure 1 shows that the largest contributions to $\overline{\mathrm{Var}(d)}$ come from the levels close to the root. In particular, for the neutral model at constant population size, the dominant contribution comes from the uppermost level, i.e. from $\mathrm{Var}(d_2)$. This measure contains the same information as the root balance ω_1, defined as the smaller of the two root branch sizes:

$$\mathrm{Var}(d_2) = \frac{1}{2}\left[\left(\omega_1 - \frac{n}{2}\right)^2 + \left(n - \omega_1 - \frac{n}{2}\right)^2\right] = \left(\frac{n}{2} - \omega_1\right)^2 \tag{5}$$

Hence the imbalance measure $\overline{\mathrm{Var}(d)}$ depends strongly on the root balance ω_1, which has been previously recognised as a meaningful global measure of tree balance (Ferretti *et al.*, 2013; Li and Wiehe, 2013), and on the imbalances of the first upper splits as well.

Estimators of θ and neutrality tests

A fundamental population genetic quantity is the scaled mutation rate $\theta = 2pN_e\mu$, where μ is the mutation rate per generation per sequence. θ is the key parameter of the neutral mutation-drift equilibrium. Usually, it cannot be measured directly, but only be estimated from observable data. For example, under the standard neutral model (*i.e.* constant population size) an unbiased estimator of θ is Watterson's $\hat{\theta}_W = S/a_n$, where S is the number of observed polymorphic sites in a sequence sample of size n ("segregating sites") (Watterson, 1975).

More generally, it has been shown that many of the well-known θ-estimators can be expressed as linear combinations of the components ξ_i of the SFS (Tajima, 1983; Achaz, 2009; Ferretti *et al.*, 2010). For example, $\hat{\theta}_W = \sum_{i=1}^{n-1} \frac{1}{a_n}\xi_i$ or Tajima's $\hat{\theta}_\pi = \sum_{i=1}^{n-1} \frac{2i(n-i)}{n(n-1)}\xi_i$. Other estimators are presented in Table 1. Furthermore, the classical neutrality tests (in their non-normalized version) can be written as a difference between two θ-estimators, hence as a linear combination of the ξ_i. For instance, the non-normalized Tajima's D (Tajima, 1989) is $\hat{\theta}_\pi - \hat{\theta}_W$, while Fay and Wu's H (Fay and Wu, 2000) is $\hat{\theta}_\pi - \hat{\theta}_H$. The most common tests are presented in Table 2.

Their expression as linear combinations of the ξ_i helps to understand discrepancies between these tests through their weight functions. For instance, from the weight functions it is immediately clear that H assigns large negative weight only to ξ_i with large i (high frequency derived alleles), while D assigns negative weight to ξ_i with small and large i (rare alleles).

For each component ξ_i of the SFS, the product $i\,\xi_i$ is an unbiased estimator of $\theta = 2pN\mu$. Hence, given weights $(w_1, ..., w_{n-1})$, the weighted linear combination

$$\hat{\theta}_{\boldsymbol{w}} = \frac{1}{\sum w_i} \sum_{i=1}^{n-1} w_i\, i\, \xi_i \tag{6}$$

is also an unbiased estimator of θ. For instance, Watterson's estimator $\hat{\theta}_W = S/a_n$ follows from setting $w_i = 1/i$ in eq (6); Tajima's estimator $\hat{\theta}_\pi$ (Tajima, 1983) is obtained by letting $w_i = (n-i)$. In fact, one can write all usual θ estimators (Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000) as linear combinations of the SFS with adequate weights (Achaz, 2009) detailed in Table 1:

$$\mathcal{T}_{\boldsymbol{\Omega}} = \frac{1}{N_{\boldsymbol{\Omega}}(S)} \sum_{i=1}^{n-1} \Omega_i\, i\, \xi_i \tag{7}$$

where $N_{\boldsymbol{\Omega}}(S) = \sqrt{\mathrm{Var}(\sum_{i=1}^{n-1} \Omega_i\, i\, \xi_i)}$. In this expression, θ is usually estimated by method of moment as $\hat{\theta} = S/a_n$, $\hat{\theta}^2 = S(S-1)/(a_n^2 + b_n)$ with $b_n = \sum_{i=1}^{n-1} 1/i^2$ (Tajima, 1989). Hence the general form of $N_{\boldsymbol{\Omega}}(S)$ is $N_{\boldsymbol{\Omega}}(S) = \sqrt{\lambda_n^{\boldsymbol{\Omega}} S + \kappa_n^{\boldsymbol{\Omega}} S(S-1)}$ with appropriate coefficients $\lambda_n^{\boldsymbol{\Omega}}, \kappa_n^{\boldsymbol{\Omega}}$ reported in Table 3 for some tests.

Decomposition of the SFS and its combinations

Here we discuss the dependence of the average spectrum $\mathrm{E}(\boldsymbol{\xi})$ on tree topology and branch lengths.

The SFS is determined by the number of mutations of size i, $1 \leq i \leq n-1$. A mutation has size i if it appears on a branch of size i. We assume that mutations occur along branches according to a homogeneous Poisson process with rate μ per unit time. Fixing a tree with respect to shape and branch lengths, we can average

over the mutation process. Denoting by E_μ the expected value for the mutation process, we obtain for the mean frequency spectrum (Fu, 1995)

$$\mathrm{E}_\mu(\xi_i|T) = \theta \sum_{k=2}^{n} k\, t_k\, P(d_k = i|T)\,, \tag{8}$$

where $P(d_k = i|T)$ is the distribution of d_k, the number of descendants of the branches of level k. These probabilities depend only on the shape of the tree T and not on waiting times. The full set of $P(d_k = i|T),\ k = 2 \ldots n$, gives actually a complete description of the tree up to permutation of the leaves.

Replacing ξ_i by their mean according to eq (8), we obtain the general expression for the mean of SFS-based θ-estimators

$$\mathrm{E}_\mu(\hat{\theta}_{\boldsymbol{w}}|T) = \frac{\theta}{\sum w_i} \sum_{i=1}^{n-1} \sum_{k=2}^{n} i\, w_i\, k\, t_k\, P(d_k = i|T) \tag{9}$$

and tests

$$\mathrm{E}_\mu(\mathcal{T}_{\boldsymbol{\Omega}}|T) = f_{\boldsymbol{\Omega}}(\theta l)\ \frac{1}{l} \sum_{i=1}^{n-1} \sum_{k=2}^{n} i\, \Omega_i\, k\, t_k\, P(d_k = i|T)\,. \tag{10}$$

where the normalisation function $f_{\boldsymbol{\Omega}}$ is defined by

$$f_{\boldsymbol{\Omega}}(\theta l) = \mathrm{E}_\mu \left[\frac{S}{N_{\boldsymbol{\Omega}}(S)} \right] \tag{11}$$

and depends on θl only, since S is a Poisson variable with parameter θl.

It is also possible to condition on S as well, obtaining

$$\mathrm{E}_\mu(\xi_i|T, S) = \frac{S}{l} \sum_{k=2}^{n} k\, t_k\, P(d_k = i|T)\,, \tag{12}$$

and

$$\mathrm{E}_\mu(\mathcal{T}_{\boldsymbol{\Omega}}|T, S) = \frac{S}{N_{\boldsymbol{\Omega}}(S)}\ \frac{1}{l} \sum_{i=1}^{n-1} \sum_{k=2}^{n} i\, \Omega_i\, k\, t_k\, P(d_k = i|T)\,. \tag{13}$$

A new subclass of neutrality tests and their decomposition

Interestingly, several common tests (and estimators) are polynomials up to second order in the frequency of mutations, hence can be written in terms of a general weight function of the form

$$\Omega_i = \alpha i + \beta + \gamma/i \tag{14}$$

with appropriate values of α, β, γ satisfying

$$\alpha \frac{n(n-1)}{2} + \beta(n-1) + \gamma a_n = 0 \text{ (or } = 1 \text{ for estimators).} \tag{15}$$

For instance, $\hat{\theta}_W$ has $\alpha = \beta = 0$ and $\gamma = 1/a_n$, while $\hat{\theta}_\pi$ has $\alpha = -2/n(n-1)$, $\beta = 2/(n-1)$ and $\gamma = 0$, hence their difference Tajima's D has $\alpha = -2/n(n-1)$, $\beta = 2/(n-1)n$ and $\gamma = -1/a_n$. Coefficients for other estimators and tests can be found in Tables 1 and 2. With this special class of weights, equation (10) becomes

$$\mathrm{E}_\mu(\mathcal{T}_{\mathbf{\Omega}}|T) = \frac{f_{\mathbf{\Omega}}(\theta l)}{l} \sum_{i=1}^{n-1} \sum_{k=2}^{n} (\alpha i^2 + \beta i + \gamma)\, k\, t_k\, P(d_k = i|T)\,. \tag{16}$$

Using $\sum_{i=1}^{n-1} iP(d_k = i|T) = \mathrm{E}(d_k) = n/k$ and $\sum_{i=1}^{n-1} i^2 P(d_k = i|T) = \mathrm{Var}(d_k) + \mathrm{E}^2(d_k)$ and exchanging the order of the sums, this becomes

$$\mathrm{E}_\mu(\mathcal{T}_{\mathbf{\Omega}}|T) = f_{\mathbf{\Omega}}(\theta l) \left(\alpha \overline{\mathrm{Var}(d)} + \sum_{k=2}^{n} \frac{t_k}{l} \left(\alpha \frac{n^2}{k} + \beta n + \gamma k \right) \right) \tag{17}$$

Results

Interpretation of neutrality tests for a single locus

We consider the application of neutrality tests to a single locus without recombination, i.e. with a given genealogy. We show that some commonly used tests statistics have a simple but rigorous interpretation in terms of tree imbalance and waiting times. The tests are summarised in Table 2 and their interpretation in Table 4. The weight of the different components is illustrated in Figure 2.

Tajima's D test statistic is the most used neutrality test. It is proportional to the difference $\hat{\theta}_\pi - \hat{\theta}_W$. If positive, indicates an excess of common alleles, if negative an excess of rare alleles.

Watterson's estimator θ_W itself has a simple interpretation. In fact, its average is $\mathrm{E}_\mu(\hat{\theta}_W) = \theta \frac{l}{a_n}$, i.e. it is proportional to the total length of the tree, divided by the mean length. As such, it is independent from the tree topology. In more practical terms, it is independent on mutation frequencies.

Using the result of section with the weights in Table 2, we can re-express the mean Tajima's D as

$$\mathrm{E}_\mu(D|T) = f_D(\theta l) \left[-\frac{2}{n(n-1)} \overline{\mathrm{Var}(d)} + \frac{1}{l} \sum_{k=2}^{n} t_k \left(\frac{2n}{(n-1)} \left(1 - \frac{1}{k}\right) - \frac{k}{a_n} \right) \right] \quad (18)$$

i.e. $\mathrm{E}_\mu(D|T)$ can be decomposed into two components: one that is a linear combination of tree lengths, independent from the topology, plus a topological component that corresponds to the measure of tree imbalance $\overline{\mathrm{Var}(d)}$ introduced before.

In qualitative terms, Tajima's D is the sum of an imbalance term with negative sign plus terms that give positive weight to the ancient waiting times and negative weight to the recent ones:

$$\boxed{D \simeq \; - \text{ tree imbalance } + \text{ length of upper branches } - \text{ length of lower branches.}}$$

Therefore, Tajima's D is large and positive when there are long branches close to the root. It is strongly negative when the tree is unbalanced and/or when recent branches are long. Tajima's D is thus sensitive to both unbalanced trees and trees with long branches close to the leafs (when negative) and balanced trees with long branches close to the root (when positive). The former are typical trees for recently increasing populations or loci under directional selection, the latter are typical under balancing selection or for structured populations or contractions in population size.

Fay and Wu's H test statistics was specifically designed to detect selective sweeps at partially linked loci, as most weight is given to derived alleles with high frequency. Strongly negative H is caused by an excess of high-frequency derived alleles, which is a signature of a locus "hitchhiking" on a nearby sweep locus (Fay

and Wu, 2000). In this paper we always consider the normalized version of this test (Zeng *et al.*, 2006). We can rewrite its mean value as

$$\mathrm{E}_\mu(H|T) = f_H(\theta l) \left[-\frac{4}{n(n-1)} \overline{\mathrm{Var}(d)} + \frac{2n}{n-1} \frac{1}{l} \sum_{k=2}^{n} t_k (1 - 2/k) \right] . \tag{19}$$

Like Tajima's D, H contains the imbalance term with negative sign. However, it has another contribution that weights positively the waiting times close to the leafs – which is opposite to Tajima's D:

$$\boxed{H \simeq \ - \text{ tree imbalance } + \text{ length of lower branches.}}$$

Therefore, H is strongly negative for (i) large imbalance, and (ii) long branches close to the root. This is precisely the signal expected by hitchhiking in the proximity of strong selective sweeps, i.e. when the sweep locus itself is uncoupled from the locus under consideration by one (or a few) recombination event(s).

Zeng's E test statistics is another test designed to detect selective sweeps. However, it is known to be less powerful than H (Zeng *et al.*, 2006). It is defined by $\hat{\theta}_L - \hat{\theta}_W$ where the estimator $\hat{\theta}_L$ has also a simple interpretation: $\mathrm{E}_\mu(\hat{\theta}_L) = \theta \frac{n}{n-1} h$ is the height h of the tree divided by the expected height. Unsurprisingly, the test is therefore a comparison of height and length of the tree:

$$\mathrm{E}_\mu(E|T) = \frac{f_E(\theta l)}{l} \left(\frac{n}{n-1} h - \frac{l}{a_n} \right) = \frac{f_E(\theta l)}{l} \sum_{k=2}^{n} \left(\frac{n}{n-1} - \frac{k}{a_n} \right) t_k \, , \tag{20}$$

$$\boxed{E \simeq \ + \text{ tree height } - \text{ tree length,}}$$

that can be rephrased as

$$\boxed{E \simeq \ + \text{ length of upper branches } - \text{ length of lower branches.}}$$

Like Fay and Wu's H, the E-test is focused on high-frequency alleles. However, it uses no topological information, but depends only on waiting times. This explains its lower power compared to other tests. Furthermore, since E compares upper and lower branches, it can actually be naturally interpreted as a test for star-likeness of a tree. In star-like trees, the length is maximal with respect to the height ($l = nh$), corresponding to strongly negative values of E.

Finally, we will discuss a common test not included in eq (14). Fu and Li's D_{FL} is one of several tests based on singletons. Its mean is $\mathrm{E}_\mu(D_{FL}|T) \propto l - \sum_{k=2}^{n} k t_k P_{n,k}(1|T)$, hence this test should measures the relative contribution of external branches to total tree length:

$$D_{FL} \simeq \ + \text{length of internal branches} \ - \ \text{length of external branches} \quad (21)$$

i.e., negative values of this test should signal extremely unbalanced (caterpillar) trees or star-like trees. However, despite its intuitive interpretation, negative values of Fu and Li's D_{FL} can be misleading if interpreted in terms of tree shapes. The reason is that these values of the test can be a result of purifying selection - non-neutral mutations that decrease fitness and therefore can only reach low frequencies before disappearing from the population. These mutations appear mostly as singletons concentrated on the lower branches. This scenario violates the assumption of mutational homogeneity along the tree and therefore the interpretation of eq (21) is not valid anymore.

A new neutrality test for positive selection

The family of tests described by eq. (14) includes Tajima's D, Fay and Wu's H and Zeng's E among many others. These three tests are built as differences of four different estimators $\hat{\theta}_W$, $\hat{\theta}_\pi$, $\hat{\theta}_L$, $\hat{\theta}_H$. However, they do not exhaust all combinations of these estimators. There is another combination[2] that has not been studied previously and will be detailed in this section.

The new test is the difference between the Watterson estimator $\hat{\theta}_W$ and the Fay and Wu's estimator $\hat{\theta}_H$. We denote this test by L:

$$L = \frac{\hat{\theta}_W - \hat{\theta}_H}{\sqrt{\mathrm{Var}\left(\hat{\theta}_W - \hat{\theta}_H\right)}} \quad (22)$$

The test compares the amount of high-frequency polymorphisms with the total number of polymorphisms.

[2]Since $\hat{\theta}_\pi = 2\hat{\theta}_L - \hat{\theta}_H$, the other two combinations $\hat{\theta}_L - \hat{\theta}_H$ and $\hat{\theta}_\pi - \hat{\theta}_L$ are equivalent to Fay and Wu's H.

The L test belongs to the family described by eq. (14), with weights $\alpha = -\frac{2}{n(n-1)}$, $\beta = 0$ and $\gamma = \frac{1}{a_n}$. Its precise definition is

$$L = \frac{\sum_{i=1}^{n-1} \left(\frac{1}{a_n} - \frac{2i^2}{n(n-1)} \right) \xi_i}{\sqrt{\lambda_n^L S + \kappa_n^L S(S-1)}} \tag{23}$$

where the coefficients λ_n^L, κ_n^L are given in Table 3. Their derivation can be found in Supplementary Material.

The interpretation of the test can be read from eq. (17):

$$\mathrm{E}_\mu(L|T) = f_L(\theta l) \left[-\frac{2}{n(n-1)} \overline{\mathrm{Var}(d)} + \frac{1}{l} \sum_{k=2}^{n} t_k \left(\frac{k}{a_n} - \frac{2n}{(n-1)} \frac{1}{k} \right) \right] \tag{24}$$

Its qualitative interpretation is different from all previous tests. It is the sum of an imbalance term with negative sign, plus negative weight to the ancient waiting times and positive weight to the recent ones:

$$\boxed{L \simeq \ - \text{ tree imbalance } - \text{ length of upper branches } + \text{ length of lower branches.}}$$

This interpretation and the presence of the Fay and Wu's estimator $\hat{\theta}_H$ in the test suggest that this test could be most powerful in selective scenarios.

In fact, simulations of the statistical power of the test in Figures 3-6 show that the left tail of L has a power similar to the normalized Fay and Wu's H test for hitchhiking (but slightly lower for most parameters). On the other hand, the right tail of L has a power similar to the left tail of Zeng's E, performing well immediately after fixation and outperforming most other tests at intermediate times after fixation. The test seems therefore to retain some of the advantages of Fay and Wu's H, while being able to detect different selective signals as well.

Extreme trees and extreme mean values of neutrality tests

Our precise interpretation of the expected values of neutrality tests in terms of tree shape and waiting times allows us to find both the extreme expected values of the tests and the corresponding "extreme" trees.

In this section we will compute the maximum and minimum value of $\mathrm{E}_\mu(\mathcal{T}_{\mathbf{\Omega}}|T,S)$, i.e. the maximum and minimum expected values of the test across all trees T, for a given number of mutations S and a given sample size n. For large S, these values depend only on the sample size. The extreme values are presented in Figure 7 as a function of n and for different values of S.

The expected value for all tests described by eq. (14) is a linear combinations of imbalances $\mathrm{Var}(d_k)$ with coefficients of the same sign:

$$\mathrm{E}_\mu(\mathcal{T}_{\mathbf{\Omega}}|T,S) = \frac{S}{N_{\mathbf{\Omega}}(S)} \sum_{k=2}^{n} \frac{kt_k}{l} \left[\alpha \mathrm{Var}(d_k) + \left(\alpha \frac{n^2}{k^2} + \beta \frac{n}{k} + \gamma\right)\right] \tag{25}$$

For this reason, maximum and minimum values correspond to maximally balanced or unbalanced topologies. Hence, to obtain these values, it is sufficient to replace $\mathrm{Var}(d_k)$ by its maximum or minimum, then maximize/minimize the result over the waiting times t_k/l (see Supplementary Information). The maximum imbalance is given by eq. (1) while the minimum imbalance will be approximated by $\min_T \mathrm{Var}(d_k) \approx 0$. In the following we will also use the related approximation $\left\lfloor \frac{n}{k} \right\rfloor \approx \frac{n}{k}$. Both approximations are correct up to $O(1/n)$ for large trees.

Tajima's D: its maximum corresponds to a tree with maximally balanced topology and length concentrated in the upmost branches ($k=2$), while its minimum corresponds to all maximally unbalanced trees with length concentrated in the upmost and lowest branches ($k=2,n$). The corresponding values are

$$\max_T \mathrm{E}_\mu(D|T,S) = \frac{\left(\frac{n}{2(n-1)} - \frac{1}{a_n}\right) S}{\sqrt{\lambda_n^D S + \kappa_n^D S(S-1)}} \underset{S\gg 1}{\longrightarrow} \frac{\frac{n}{2(n-1)} - \frac{1}{a_n}}{\sqrt{\kappa_n^D}} \underset{n\gg 1}{\longrightarrow} \frac{3}{2\sqrt{2}} \log(n) \tag{26}$$

$$\min_T \mathrm{E}_\mu(D|T,S) = \frac{\left(\frac{2}{n} - \frac{1}{a_n}\right) S}{\sqrt{\lambda_n^D S + \kappa_n^D S(S-1)}} \underset{S\gg 1}{\longrightarrow} \frac{\frac{2}{n} - \frac{1}{a_n}}{\sqrt{\kappa_n^D}} \underset{n\gg 1}{\longrightarrow} -\frac{3}{\sqrt{2}} \approx -2.1 \tag{27}$$

where the first arrow in each equation represents the limit of large number of segregating sites, and the second the asymptotic behaviour for large sample size. The maximum and minimum values of $\mathrm{E}_\mu(D|T,S)$ are also the absolute maximum and minimum values of D over all possible spectra.

Fay and Wu's H: its maximum corresponds to a tree with maximally balanced topology and length concentrated (surprisingly) in branches at $k = 4$, while its minimum corresponds to a maximally unbalanced tree with length concentrated in the upmost branches ($k = 2$). The corresponding values are

$$\max_T \mathrm{E}_\mu(H|T,S) = \frac{\frac{n}{4(n-1)}S}{\sqrt{\lambda_n^H S + \kappa_n^H S(S-1)}} \underset{S\gg 1}{\longrightarrow} \frac{\frac{n}{4(n-1)}}{\sqrt{\kappa_n^H}} \underset{n\gg 1}{\longrightarrow} \frac{\log(n)}{4\sqrt{\pi^2 - 88/9}} \tag{28}$$

$$\min_T \mathrm{E}_\mu(H|T,S) = \frac{-\frac{(n-2)^2}{n(n-1)}S}{\sqrt{\lambda_n^H S + \kappa_n^H S(S-1)}} \underset{S\gg 1}{\longrightarrow} -\frac{\frac{(n-2)^2}{n(n-1)}}{\sqrt{\kappa_n^H}} \underset{n\gg 1}{\longrightarrow} -\frac{\log(n)}{\sqrt{\pi^2 - 88/9}} \tag{29}$$

Zeng's E: its maximum corresponds to a tree with length concentrated in the upper branches ($k = 2$), while its minimum corresponds to star-like trees (i.e. length concentrated in the lowest branches $k = n$). The corresponding values are

$$\max_T \mathrm{E}_\mu(E|T,S) = \frac{\left(\frac{n}{2(n-1)} - \frac{1}{a_n}\right)S}{\sqrt{\lambda_n^E S + \kappa_n^E S(S-1)}} \underset{S\gg 1}{\longrightarrow} \frac{\frac{n}{2(n-1)} - \frac{1}{a_n}}{\sqrt{\kappa_n^E}} \underset{n\gg 1}{\longrightarrow} \frac{1}{2}\sqrt{\frac{3}{\pi^2 - 9}}\log(n) \tag{30}$$

$$\min_T \mathrm{E}_\mu(E|T,S) = \frac{\left(\frac{1}{n-1} - \frac{1}{a_n}\right)S}{\sqrt{\lambda_n^E S + \kappa_n^E S(S-1)}} \underset{S\gg 1}{\longrightarrow} \frac{\frac{1}{n-1} - \frac{1}{a_n}}{\sqrt{\kappa_n^E}} \underset{n\gg 1}{\longrightarrow} -\sqrt{\frac{3}{\pi^2 - 9}} \approx -1.9 \tag{31}$$

The minimum value of $\mathrm{E}_\mu(E|T,S)$ is also the minimum absolute values of E.

L **test:** its maximum corresponds to a star-like tree with length concentrated in the lowest branches ($k = n$), while its minimum corresponds to a maximally unbalanced tree with length concentrated in the upmost branches ($k = 2$). The corresponding values are

$$\max_T \mathrm{E}_\mu(L|T,S) = \frac{\left(\frac{1}{a_n} - \frac{2}{n(n-1)}\right)S}{\sqrt{\lambda_n^L S + \kappa_n^L S(S-1)}} \underset{S\gg 1}{\longrightarrow} \frac{\frac{1}{a_n} - \frac{2}{n(n-1)}}{\sqrt{\kappa_n^L}} \underset{n\gg 1}{\longrightarrow} \frac{3}{2\sqrt{6\pi^2 - 29}} \approx 0.3 \tag{32}$$

$$\min_T \mathrm{E}_\mu(L|T,S) = \frac{\left(\frac{1}{a_n} - \frac{(n-1)^2+1}{n(n-1)}\right)S}{\sqrt{\lambda_n^L S + \kappa_n^L S(S-1)}} \underset{S\gg 1}{\longrightarrow} \frac{\frac{1}{a_n} - \frac{(n-1)^2+1}{n(n-1)}}{\sqrt{\kappa_n^L}} \underset{n\gg 1}{\longrightarrow} -\frac{3}{2\sqrt{6\pi^2 - 29}}\log(n) \tag{33}$$

The maximum value of $\mathrm{E}_\mu(L|T,S)$ is also the maximum absolute values of L.

The dependence of the extreme values on n and S is shown in Figure 7 for all the tests discussed above.

These results are useful to interpret the actual strength of the signal given by the tests. The normalisation of neutrality tests suggests that values between -1 and 1 fall into the normal range for realizations of the neutral model without recombination. However, there is no indication of which values could be deemed "large" in absolute terms. The extreme values computed above fill this gap, since they give a natural reference in terms of extreme trees. These values can be used to see if a tree is close to one of the extreme trees for the test used, and to understand how large a signal of non-neutrality could be in theory.

As an example, consider the regions of the human genome shown in Figure 8. The strong signals of selection in Central Europeans detected by Fay and Wu's H appear much less extreme when compared with the theoretical minimum, which is so low that it does not appear in the plot. On the other hand, the deviations from neutrality shown by Tajima's D around 136.4Mbp of chromosome 2 and around 29.4Mbp and 30.6Mbp of chromosome 6 in Central Europeans do not look impressive, unless we notice that it is pretty close to the minimum possible value for the test. As another example, the deviations from neutrality of L and Fay and Wu's H around 31.3Mbp on chromosome 6 in Yoruba look similar, but the minimum of L is much closer.

The results of this section could also be used to renormalise neutrality tests in the spirit of Schaeffer (2002) (see Supplementary Information). However, our results could not actually show any improvement with respect to the usual normalisation.

Discussion and conclusions

The ancestry of the sequences in a sample from a single locus, or an asexual population, is described by a single genealogical tree. The same is not true for multi-locus analyses of sexual species: recombination generates different trees along the

genome. Inferring these trees is possible only if there are enough mutations per branch. However, in most sexual and asexual populations, lower branches are typically short compared to the inverse mutation rate. Moreover, in many eukaryotic genomes, the mutation and recombination rates are of the same order of magnitude, which means that there are just a few segregating sites in each non-recombining fragment of the genome. The paucity of mutations, caused by the interplay of genetic relatedness within a population (hence short branches) and recombination, does not allow a full reconstruction of the trees. Therefore, summary statistics are often used for population genetics analysis. These statistics are also computationally useful, since any given configuration of mutations has low probability and it is therefore hard to apply inference methods on the configuration itself. Moreover, they are more robust to details of the model like mutation and recombination rates.

Summary statistics are often more directly related to the mutation pattern of the sequences rather than to their genealogy. In this work, we clarified the precise correspondence between some SFS summary statistics and some features of the genealogical trees.

It is well known that the frequency spectrum is sensitive to tree topology and branch lengths. Interestingly, several estimators and neutrality tests built on the SFS – such as Watterson θ_W, Tajima's D, Fay and Wu's H – show a quite simple dependence on tree imbalance and waiting times. A new measure of tree imbalance – the variance in the number of descendants of a mutation at a given level – plays an important role in the interpretation of these neutrality tests. The simplicity of these results stems from the simple weights of these estimators and tests: the SFS is multiplied by functions of the frequency that are constant (Watterson), linear (Zeng) or quadratic polynomials (Fay and Wu, Tajima).

The interpretation of common estimators and tests is summarised in Table 4. This interpretation is rigorous and consistent with intuition. Our results help to understand the peculiarities of the different tests. For example, we re-interpret Zeng's E as a test for star-likeness, and understand its reduced power to detect selection compared to Fay and Wu's H as a consequence of its insensitivity to tree

imbalance and of the compressed distributions of its negative values.

The imbalance measure $\overline{\mathrm{Var}(d)}$ is also related to other balance statistics proposed recently, namely the root balance ω_1 and the standardized sum $\omega_1 + \omega_2 + \omega_3$ (Li and Wiehe, 2013), which can also be inferred quite reliably from sequence data. In contrast, balance statistics such as Colless' index (Colless, 1982), which considers the average balance of the tree across all internal nodes, are less suited for population genetic applications, since balance at lower nodes can usually not be estimated from sequence data, due to the paucity of polymorphisms which separate closely related sequences. Furthermore, recombination affects mostly the lower part of the tree, hence it introduces additional noise preventing accurate reconstruction of its topology. Further studies of $\overline{\mathrm{Var}(d)}$ and similar imbalance measures on phylodynamic trees could provide some interesting summary statistics.

The limitation of the approach presented here lies in the assumption that mutations are mostly neutral and the mutation rate is constant, *i.e.* mutations should occur randomly on the tree. This assumption fails for the case of purifying selection, when deleterious mutations can be more abundant than neutral ones and tend to accumulate on the lower branches of the tree. In fact, for sequences under purifying selection, the topology of the tree itself depends on the deleterious mutations. Therefore our approach could not work for tests aimed at detecting rare alleles under purifying selection, like Fu and Li's tests (or extreme negative values of Tajima's D).

Beyond clarifying the interpretation of existing tests, our results open some possibilities for building new neutrality tests to explore different aspects of tree shape. Our new L test is a simple test for selection that shows an interesting behaviour, with power similar to Fay and Wu's H in left tail and to Zeng's E in the right tail, and therefore is able to detect deviations from neutrality in hitchhiking and selective scenarios at different times and different recombination rates. This new L test is in the same class as Tajima's D and the other tests, hence it is sensitive to the variance $\overline{\mathrm{Var}(d)}$. New tests in the same class are possible, but one could imagine other tests sensitive e.g. to different combinations of the variances $\mathrm{Var}(d_k)$ or to the skewness or kurtosis of $P(d_k = i|T)$ as well. While the variance

$\overline{\mathrm{Var}(d)}$ is a direct measure of imbalance and especially to the imbalance of the upper branches, other combinations could be sensitive to different tree features.

While our results help to interpret positive and negative values of the tests, they also provide information about the size of these values. Given the normalisation of the tests, it is well known that the typical range of values of the standard neutral model is ± 1, and confidence intervals can be computed by coalescent simulations, but this says nothing about the size of deviations from this model. Our results on extreme trees and the corresponding extreme test values give some indication on the range of potential deviations from neutrality.

Finally, our approach can be used to understand the average structure of the genealogical trees generated by models for which the expected SFS is known. Some of our results could also find application in phylogenetic studies of closely related species or populations, where the reconstruction of the phylogenetic tree could be difficult or ambiguous.

Acknowlegments

This work was stimulated by discussions with Michael Blum and Filippo Disanto. We thank an anonymous reviewer for useful comments. AL is funded by the UK National Institute for Health Research (NIHR) Health Protection Research Unit on Modelling Methodology (grant HPRU-2012-10080). LF and GA acknowledge support from the grant ANR-12-JSV7-0007 from Agence Nationale de Recherche (France). GA acknowledges support from the grant ANR-12-BSV7-0012-04 from Agence Nationale de Recherche (France). TW acknowledges support from DFG-SPP1590 by the German Science Foundation.

Estimator	formula	weights w_i	α	β	γ	reference
$\hat{\theta}_W$	$\frac{\sum_{i=1}^{n-1}\xi_i}{a_n}$	$1/ia_n$	0	0	$\frac{1}{a_n}$	Watterson (1975)
$\hat{\theta}_\pi$	$\frac{2\sum_{i=1}^{n-1} i(n-i)\xi_i}{n(n-1)}$	$(n-i)/\binom{n}{2}$	$-\frac{2}{n(n-1)}$	$\frac{2}{n-1}$	0	Tajima (1983)
$\hat{\theta}_L$	$\frac{\sum_{i=1}^{n-1} i\xi_i}{n-1}$	$1/(n-1)$	0	$\frac{1}{n-1}$	0	Zeng *et al.* (2006)
$\hat{\theta}_H$	$\frac{2\sum_{i=1}^{n-1} i^2\xi_i}{n(n-1)}$	$i/\binom{n}{2}$	$\frac{2}{n(n-1)}$	0	0	Fay and Wu (2000)
$\hat{\theta}_{\xi_1}$	ξ_1	$\delta_{i,1}$	-	-	-	Fu and Li (1993)

Table 1: Selected unbiased linear estimators of θ.

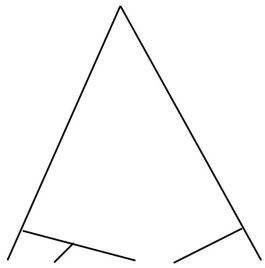
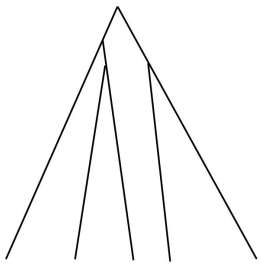
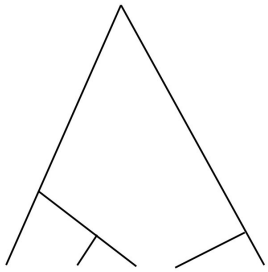
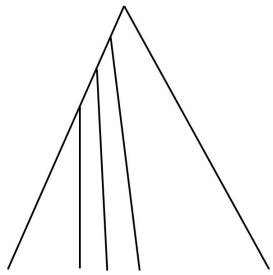
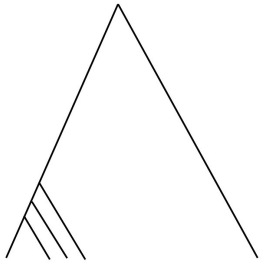
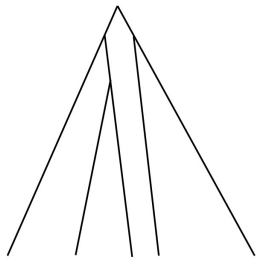
Test	formula	weights Ω_i	α	β	γ	reference
D	$\hat{\theta}_\pi - \hat{\theta}_W$	$(n-i)/\binom{n}{2} - 1/ia_n$	$-\frac{2}{n(n-1)}$	$\frac{2}{n-1}$	$-\frac{1}{a_n}$	Tajima (1989)
H	$\hat{\theta}_\pi - \hat{\theta}_H$	$(n-2i)/\binom{n}{2}$	$-\frac{4}{n(n-1)}$	$\frac{2}{n-1}$	0	Fay and Wu (2000)
E	$\hat{\theta}_L - \hat{\theta}_W$	$1/(n-1) - 1/ia_n$	0	$\frac{1}{n-1}$	$-\frac{1}{a_n}$	Zeng *et al.* (2006)
L	$\hat{\theta}_W - \hat{\theta}_H$	$1/ia_n - i/\binom{n}{2}$	$-\frac{2}{n(n-1)}$	0	$\frac{1}{a_n}$	this study
D_{FL}	$\hat{\theta}_{\xi_1} - \hat{\theta}_W$	$\delta_{i,1} - 1/ia_n$	-	-	-	Fu and Li (1993)

Table 2: Neutrality tests discussed in this paper.

Test	$\lambda_n^{\mathbf{\Omega}}$	$\kappa_n^{\mathbf{\Omega}}$
D	$\frac{n+1}{3(n-1)a_n} - \frac{1}{a_n^2}$	$\frac{1}{a_n^2+b_n}\left[\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n-2}{na_n} + \frac{b_n}{a_n^2}\right]$
H	$\frac{n-2}{6(n-1)a_n}$	$\frac{18n^2(3n+2)b_{n+1}-(88n^3+9n^2-13n+6)}{9n(n-1)^2(a_n^2+b_n)}$
E	$\frac{n}{2(n-1)a_n} - \frac{1}{a_n^2}$	$\frac{1}{a_n^2+b_n}\left[\frac{b_n}{a_n^2} + 2\left(\frac{n}{n-1}\right)^2 b_n - \frac{2(nb_n-n+1)}{(n-1)a_n} - \frac{3n+1}{n-1}\right]$
L	$\frac{1}{a_n}\left(1 - \frac{1}{a_n}\right)$	$\frac{1}{a_n^2+b_n}\left[\frac{b_n}{a_n^2} + 2\frac{36n^2(2n+1)b_{n+1}-116n^3+9n^2+2n-3}{9n(n-1)^2} - \frac{4}{n(n-1)a_n}\left(n^2 b_n - \frac{(5n+2)(n-1)}{4}\right)\right]$

Table 3: Coefficients of the normalisation of the neutrality tests discussed in this paper.

Table 4: Interpreting neutrality tests

Test:	**Tajima's** D	**Fay and Wu's** H	**Zeng's** E
Spectrum:	common vs rare alleles	common vs high-frequency alleles	high-frequency vs low-frequency alleles
Interpretation:	- tree imbalance + length of upper branches - length of lower branches	- tree imbalance + length of lower branches	height - length (= length of upper branches - length of lower branches)
Tree: test > 0	population structure: balanced tree, long root branches	balanced tree, starlike	long root branches
Example: test > 0			
Tree: test < 0	starlike or unbalanced tree	hitchhiking: unbalanced tree, long root branches	starlike
Example: test < 0			

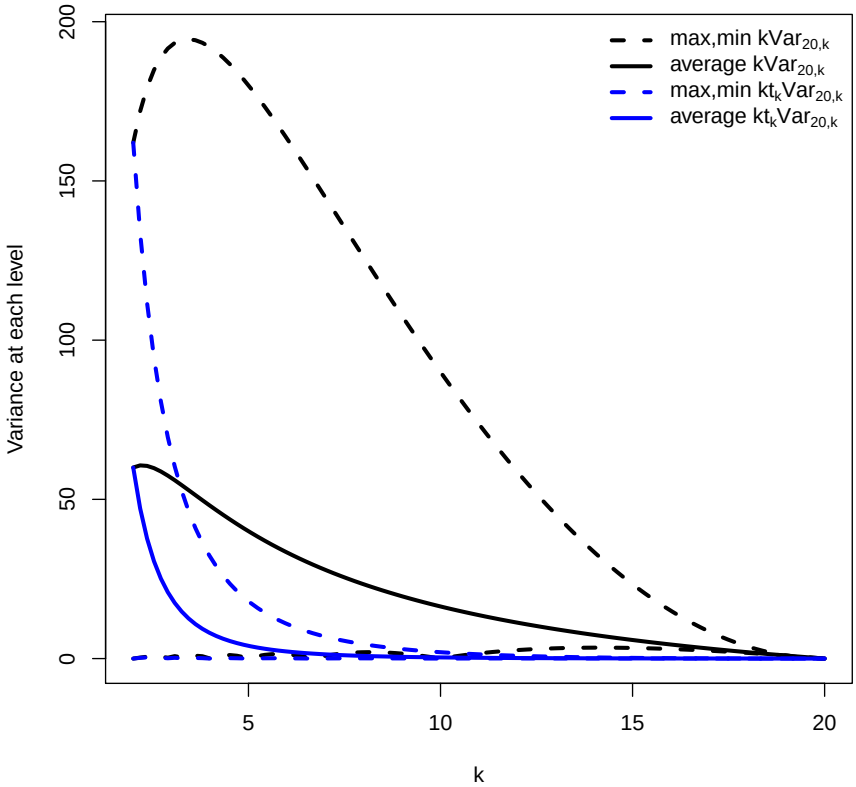


Figure 1: Plot of the mean, maximum and minimum contributions of different levels $k = 2 \ldots 20$ to the variance $\overline{\mathrm{Var}(d)}l$, for a sample with $n = 20$. In black the contribution per unit waiting time; in red, the total contribution per level in the Kingman coalescent.

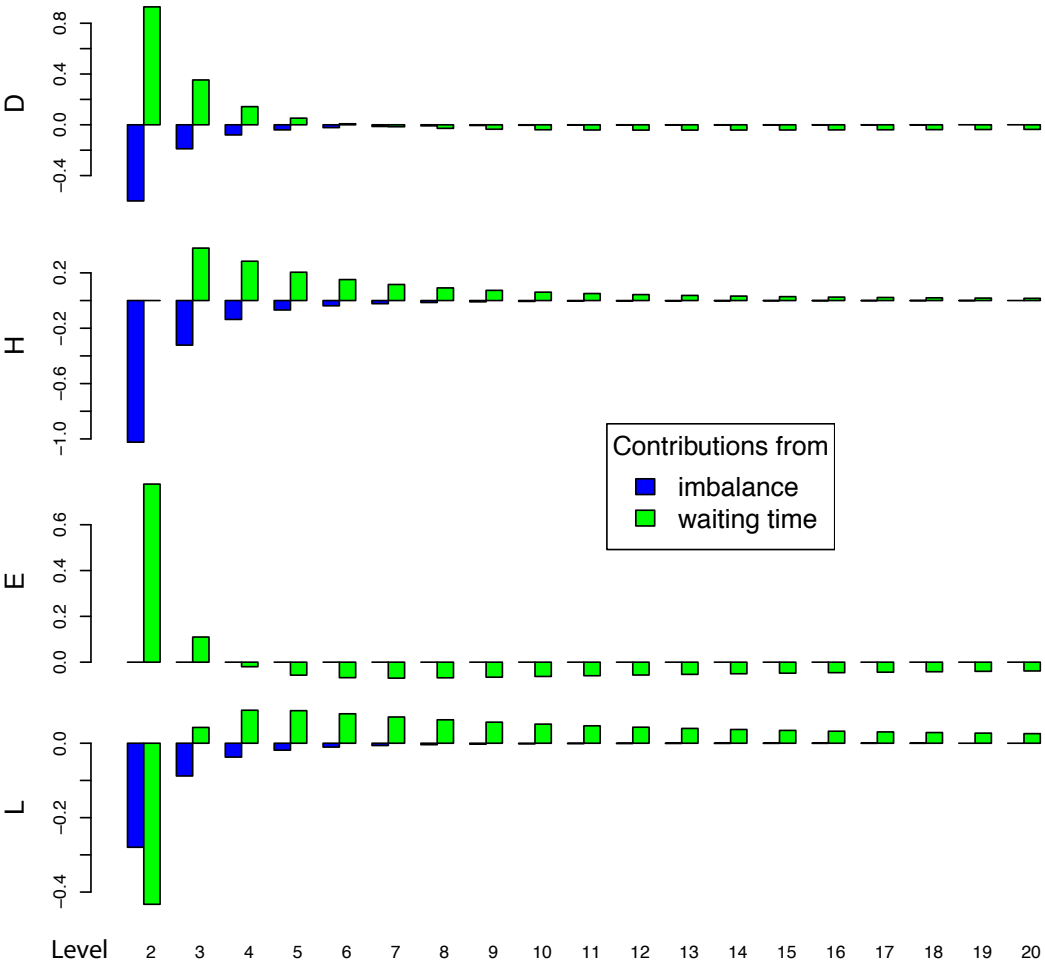


Figure 2: Plot of the mean contribution to the value of the tests of each imbalance component $\mathrm{Var}(d_k)$ (blue) and each residual purely waiting time component t_k (green) under neutrality (i.e. for the Kingman coalescent). The sum of all contributions for each test is zero. Contributions are shown for different levels $k = 2 \ldots 20$ in a sample with $n = 20$ individuals and $S = 20$.

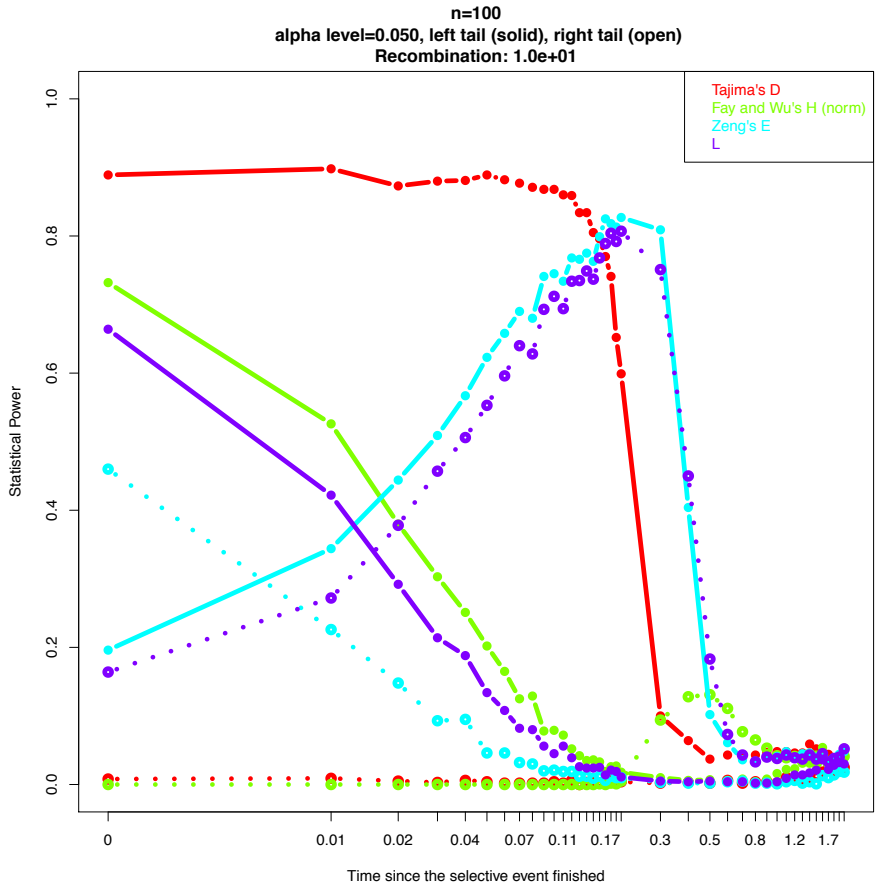


Figure 3: Statistical power of L and other neutrality tests to detect hitchhiking against the standard neutral model. Coalescent simulations performed with *mstatspop* (Ramos-Onsins) for a sample of size $n = 100$ in a population of size $N_e = 10^6$, for sequences of length 10^5 bp and $\theta = 10^{-3}$/bp, located 1 Mbp away from a selected sites with selection coefficient $4N_e s = 10^3$. Recombination rate $4N_e r = 10$ with respect to the selected site.

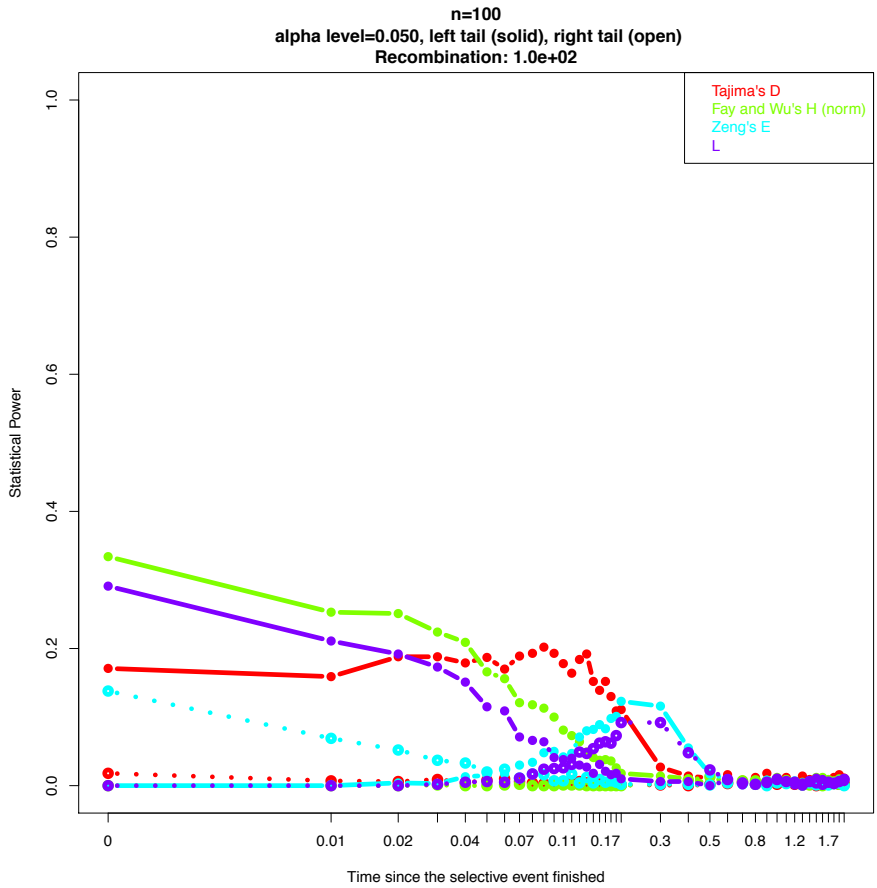


Figure 4: Statistical power of L and other neutrality tests to detect hitchhiking against the standard neutral model. Recombination rate $4Nr = 100$ with respect to the selected site.

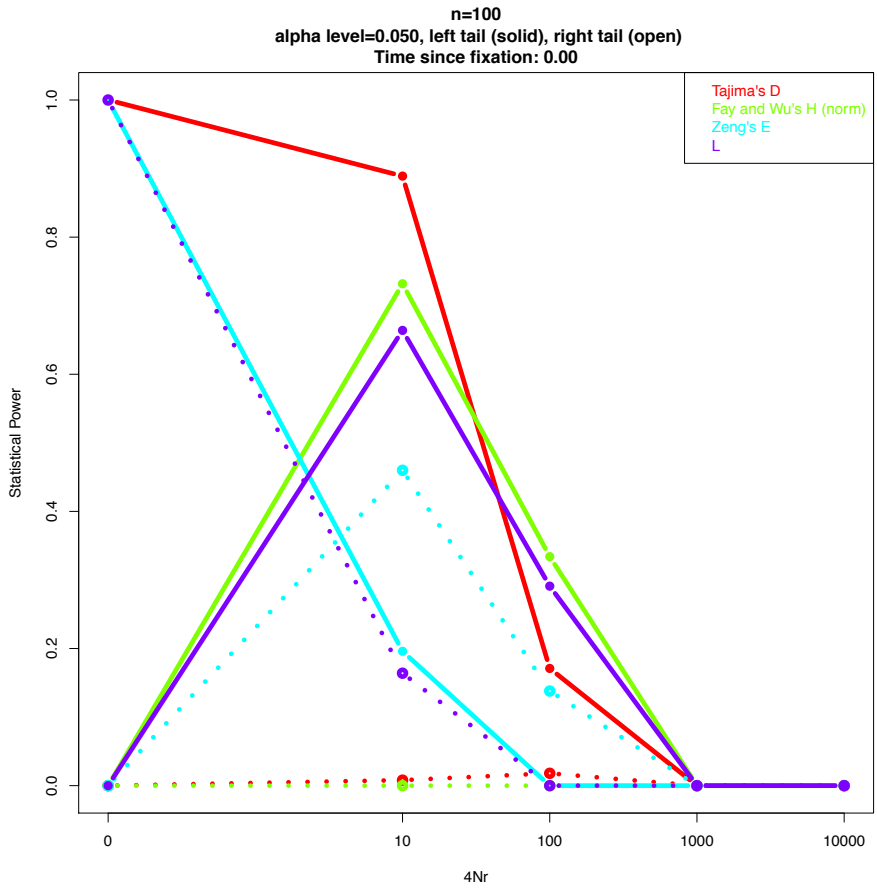


Figure 5: Statistical power of L and other neutrality tests to detect hitchhiking against the standard neutral model, immediately after fixation of the selected allele.

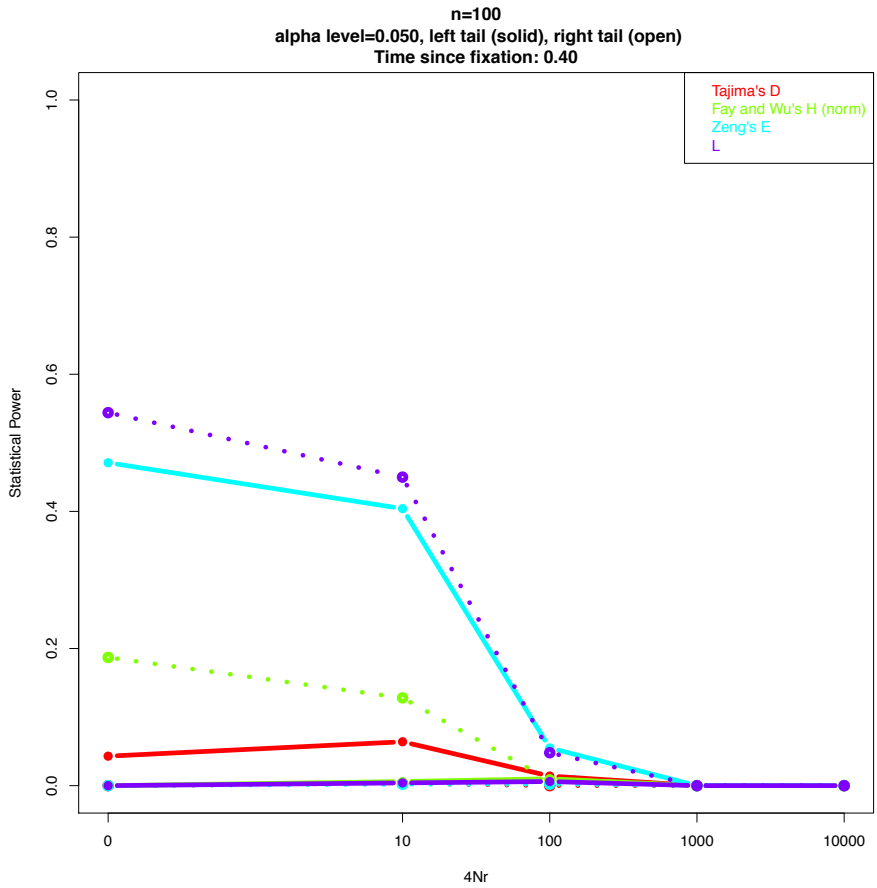


Figure 6: Statistical power of L and other neutrality tests to detect hitchhiking against the standard neutral model, 0.4 coalescent times after fixation of the selected allele.

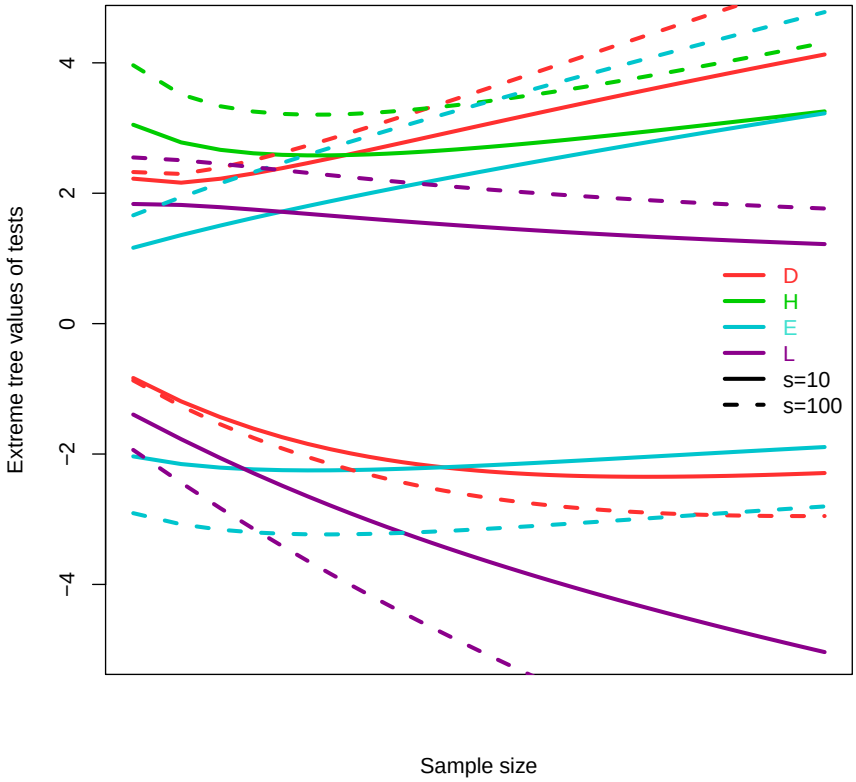


Figure 7: Maximum and minimum values of neutrality tests as a function of n for $S = 10, 100$. The minimum of Fay and Wu's H is not shown since its decreases from about -10 to -30 in the range of sample sizes of the plot.

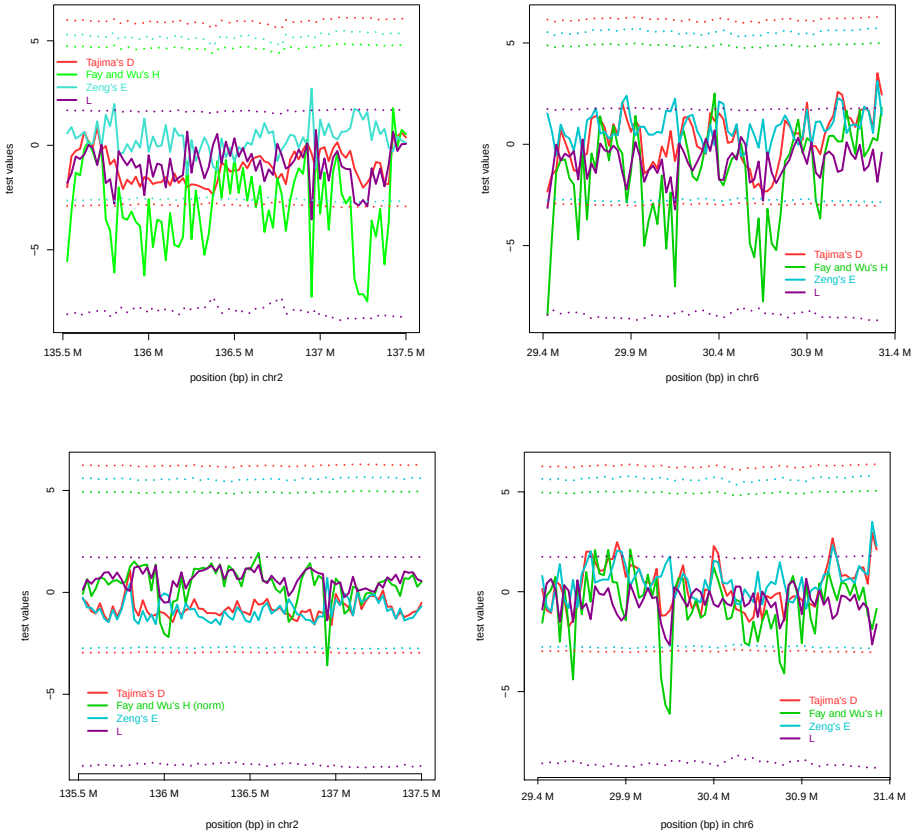


Figure 8: Values of neutrality tests, compared to their extreme values (dotted lines), in the region surrounding the LCT (left) and MHC gene (right) in human. The values are computed from 1000 Genomes Project data (1000 Genomes Project Consortium *et al.*, 2015) for about 100 diploid individuals from Central European (above) and Yoruba populations (below) in windows of 25 kb. (The minimum of Fay and Wu's H lies around -30 for all plots.)

Literature Cited

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

Achaz, G., 2009 Frequency Spectrum Neutrality Tests: One for All and All for One. Genetics **183**: 249.

Blum, M. G., and O. François, 2005 On statistical tests of phylogenetic tree imbalance: the sackin and other indices revisited. Mathematical biosciences **195**: 141–153.

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

Blum, M. G., O. François, and S. Janson, 2006 The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. The Annals of Applied Probability : 2195–2214.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, *et al.*, 2014 Beast 2: a software platform for bayesian evolutionary analysis. PLoS Comput Biol **10**: e1003537.

Colless, D., 1982 Review of phylogenetics: the theory and practice of phylogenetic systematics. Syst. Zool **31**: 100–104.

Fay, J., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155**: 1405.

Felsenstein, J., 2004 Inferring phylogenies .

Ferretti, L., F. Disanto, and T. Wiehe, 2013 The effect of single recombination events on coalescent tree height and shape. PloS one **8**: e60123.

Ferretti, L., M. Perez-Enciso, and S. Ramos-Onsins, 2010 Optimal neutrality tests based on the frequency spectrum. Genetics **186**: 353–365.

Fu, Y., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693.

Fu, Y.-X., 1995 Statistical properties of segregating sites. Theoretical Population Biology **48**: 172–197.

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

Hein, J., M. Schierup, and C. Wiuf, 2004 *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford university press.

Ho, S. Y., and B. Shapiro, 2011 Skyline-plot methods for estimating demographic history from nucleotide sequences. Molecular Ecology Resources **11**: 423–434.

Kimura, M., 1985 *The neutral theory of molecular evolution*. Cambridge University Press.

Kingman, J. F., 1982 On the genealogy of large populations. Journal of Applied Probability : 27–43.

Lapierre, M., C. Blin, A. Lambert, G. Achaz, and E. P. Rocha, 2016 The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. Molecular biology and evolution : msw048.

Li, H., and T. Wiehe, 2013 Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation. PLoS Computational Biology **9**.

Liu, X., and Y.-X. Fu, 2015 Exploring population size changes using snp frequency spectra. Nature genetics .

Pybus, O. G., A. Rambaut, and P. H. Harvey, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155**: 1429–1437.

Schaeffer, S. W., 2002 Molecular population genetics of sequence length diversity in the adh region of drosophila pseudoobscura. Genetical research **80**: 163–175.

Sloane, N., and S. Plouffe, 1995 The encyclopedia of integer sequences.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585.

Wakeley, J., 2009 *Coalescent theory: an introduction*, volume 1. Roberts & Company Publishers Greenwood Village, Colorado.

Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical population biology **7**: 256.

Yule, G. U., 1925 A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character : 21–87.

Zeng, K., Y.-X. Fu, S. Shi, and C.-I. Wu, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics **174**: 1431–1439.

Zivkovic, D., and T. Wiehe, 2008 Second-Order Moments of Segregating Sites Under Variable Population Size. Genetics **180**: 341.

Supplementary Information

Derivation of the normalisation of L

The normalisation of the L test

$$\mathrm{Var}(\hat{\theta}_W - \hat{\theta}_H) = \mathrm{Var}(\hat{\theta}_W) + \mathrm{Var}(\hat{\theta}_H) - 2\mathrm{Cov}(\hat{\theta}_W, \hat{\theta}_H) \tag{S1}$$

can be derived from the known variances of the Watterson estimators (Tajima, 1983)

$$\mathrm{Var}(\hat{\theta}_W) = \frac{1}{a_n}\theta + \frac{b_n}{a_n^2}\theta^2 \tag{S2}$$

and of the Fay and Wu's estimator (Zeng *et al.*, 2006)

$$\mathrm{Var}(\hat{\theta}_H) = \theta + 2\frac{36n^2(2n+1)b_{n+1} - 116n^3 + 9n^2 + 2n - 3}{9n(n-1)^2}\theta^2 \tag{S3}$$

and from the covariance of the two estimators, that can be obtained using the results of Fu (1995) as

$$\mathrm{Cov}(\hat{\theta}_W, \hat{\theta}_H) = \sum_{i=1}^{n-1} \frac{1}{a_n}\frac{2i^2}{n(n-1)}\frac{\theta}{i} + \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} \frac{1}{a_n}\frac{2i^2}{n(n-1)}\theta^2\sigma_{ij} \tag{S4}$$

Substituting σ_{ij} with its definition from Fu (1995) and solving the sums using the combinatorial results below

$$\sum_{j=1}^{n-1} \sigma_{ij} = \frac{a_n - a_i}{n-i} \tag{S5}$$

$$\sum_{i=1}^{n-1} f(i) = \sum_{i=1}^{n-1} f(n-i) \quad \text{for any function } f \tag{S6}$$

$$\sum_{i=1}^{n-1} a_i = (n-1)(a_n - 1) \tag{S7}$$

$$\sum_{i=1}^{n-1} i a_i = \frac{n(n-1)}{2}a_n - \frac{n(n+1)-2}{4} \tag{S8}$$

we finally obtain

$$\mathrm{Cov}(\hat{\theta}_W, \hat{\theta}_H) = \frac{1}{a_n}\theta + \frac{2}{n(n-1)a_n}\left(n^2 b_n - \frac{(5n+2)(n-1)}{4}\right)\theta^2 \tag{S9}$$

Substituting the estimates S/a_n for θ and $S(S-1)/(a_n^2+b_n)$ for θ^2, we find the coefficients

$$\lambda_n^L = \frac{1}{a_n}\left(1 - \frac{1}{a_n}\right) \tag{S10}$$

$$\begin{aligned}\kappa_n^L =& \frac{1}{a_n^2+b_n}\left[\frac{b_n}{a_n^2} + 2\frac{36n^2(2n+1)b_{n+1} - 116n^3 + 9n^2 + 2n - 3}{9n(n-1)^2}\right.\\ &\left. - \frac{4}{n(n-1)a_n}\left(n^2 b_n - \frac{(5n+2)(n-1)}{4}\right)\right]\end{aligned} \tag{S11}$$

Derivation of extreme trees

The derivation for extreme trees proceeds in a different way depending if $\alpha > 0$ or $\alpha < 0$. Here we consider the case $\alpha \leq 0$ which includes all tests discussed in this paper.

For $\alpha \leq 0$, the tree imbalance affects negatively the test. Hence, the maximum of the test corresponds to maximally balanced trees (minimum $\overline{\mathrm{Var}(d)} \approx 0$).

The waiting times of the extreme tree corresponding to the maximum value can be obtained by the maximisation of the sum $\sum_{k=2}^{n} \frac{kt_k}{l}\left(\alpha\frac{n^2}{k^2} + \beta\frac{n}{k} + \gamma\right)$ over the t_ks with constraint $\sum_{k=2}^{n} kt_k = l$. If we denote

$$k_{max} = \mathrm{argmax}_{k\in[2,n]}\left(\alpha\frac{n^2}{k^2} + \beta\frac{n}{k} + \gamma\right) \tag{S12}$$

the sum discussed before is clearly maximised by

$$t_{k_{max}} = \frac{l}{k_{max}} \quad , \quad t_k = 0 \text{ for } k \neq k_{max} \tag{S13}$$

To find k_{max}, we consider k as a real variable and we find the condition for the maximum as the zero of the derivative

$$-2\alpha\frac{n^2}{k^3} - \beta\frac{n}{k^2} = 0 \quad \Rightarrow \quad k = -\frac{2\alpha n}{\beta} \tag{S14}$$

and since the derivative is positive for $k < -\frac{2\alpha n}{\beta}$ and negative for $k = -\frac{2\alpha n}{\beta}$, then k_{max} is one of the two integers closest to $-\frac{2\alpha n}{\beta}$. The two values can be compared for any given test to find k_{max}.

On the other hand, the minimum of the test corresponds to maximally imbalanced trees, i.e. maximum variance $\overline{\mathrm{Var}(d)} = (k-1)(n/k-1)^2$. Therefore, the waiting times can be obtained by minimising the sum $\sum_{k=2}^{n} \frac{kt_k}{l} \left(\alpha(k-1)\left(\frac{n}{k}-1\right)^2 + \alpha\frac{n^2}{k^2} + \beta\frac{n}{k} + \gamma\right)$ over the t_ks. If we denote

$$k_{min} = \mathrm{argmin}_{k\in[2,n]} \left(\alpha(k-1)\left(\frac{n}{k}-1\right)^2 + \alpha\frac{n^2}{k^2} + \beta\frac{n}{k} + \gamma\right) \tag{S15}$$

the sum discussed before is clearly minimised by

$$t_{k_{min}} = \frac{l}{k_{min}} \quad , \quad t_k = 0 \text{ for } k \neq k_{min} \tag{S16}$$

To find k_{min}, we consider k as a real variable and we find the condition for the minimum. First, we study the zeros of the derivative of the above sum

$$-\frac{\alpha n^2 + 2\alpha n + \beta n}{k^2} + \alpha = 0 \tag{S17}$$

that corresponds to $k = \sqrt{n(n+2+\beta/\alpha)}$. This value is a maximum and the sum is convex for positive k, hence the minimum is at one of the boundaries:

$$k_{min} = 2 \quad \text{or} \quad k_{min} = n \tag{S18}$$

The two values can be compared for any given test to find k_{min}.

Normalizing tests by their extreme values

The usual normalisation of neutrality tests does not make the results easily comparable among samples with different number of individuals or among regions with different variability/number of SNPs, because of the dependence of the values on n and S. On the other hand, it has been suggested in the past (Schaeffer, 2002) that normalising the tests by their extreme values could make it easier to compare and interpret them in terms of tree shapes, although it becomes more

difficult to grasp the significance of the test values without a full computation of the confidence intervals. A possible renormalisation would be the following:

$$\mathcal{T}'_{\mathbf{\Omega}} = \begin{cases} \frac{\mathcal{T}_{\mathbf{\Omega}}}{|\max_T \mathrm{E}_\mu(\mathcal{T}_{\mathbf{\Omega}}|T,S)|} & , \quad \mathcal{T}_{\mathbf{\Omega}} \geq 0 \\ \frac{\mathcal{T}_{\mathbf{\Omega}}}{|\min_T \mathrm{E}_\mu(\mathcal{T}_{\mathbf{\Omega}}|T,S)|} & , \quad \mathcal{T}_{\mathbf{\Omega}} < 0 \end{cases} \tag{S19}$$

For example, for Tajima's D the result would be

$$D' = \begin{cases} \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\left(\frac{n}{2(n-1)} - \frac{1}{a_n}\right)S} & , \quad \hat{\theta}_\pi - \hat{\theta}_W \geq 0 \\ \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\left(\frac{1}{a_n} - \frac{2}{n}\right)S} & , \quad \hat{\theta}_\pi - \hat{\theta}_W < 0 \end{cases} \tag{S20}$$

These newly normalized tests would show values close to 1 or -1 for trees close to the extreme ones. These tests do not depend on the absolute value of the spectrum, but only on the normalized spectrum ξ_k/S. This means that they do not depend on the number of SNPs, but only on their frequency distribution. Moreover, if the confidence intervals of the usual tests are computed by conditioning on n and S, as it is often the case, then the confidence intervals of the renormalised tests are simply the renormalised confidence intervals.

However, empirical evidence from analysis of real data suggests that this renormalisation does not make the test values more comparable among different samples with different values of n and S.

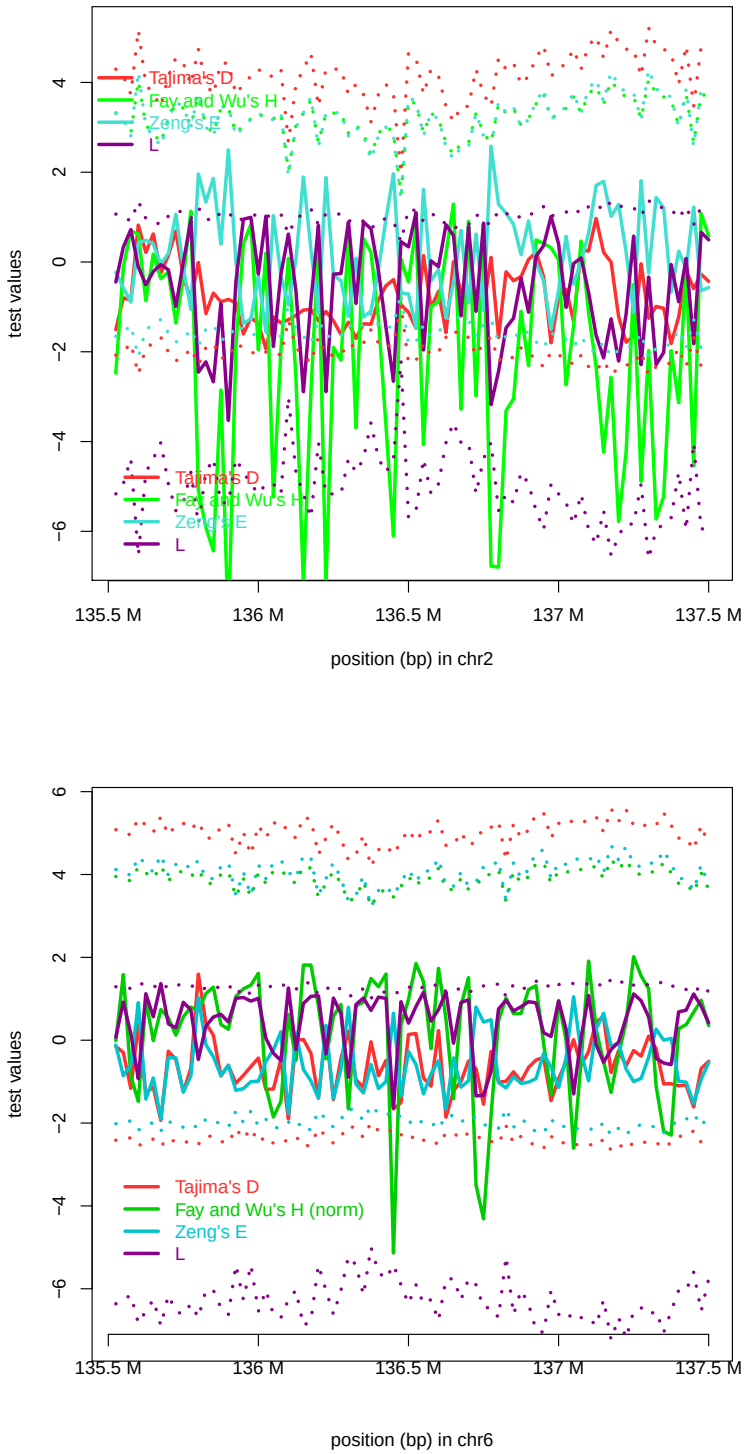


Figure S1: Values of neutrality tests, compared to their extreme values (dotted lines), in the region surrounding the LCT (above) and MHC gene (below) in human. The values are computed from 1000 Genomes Project data (1000 Genomes Project Consortium *et al.*, 2015) for about 100 diploid individuals from Central European (above) and Yoruba populations (below) in windows of 25 kb, but selecting only 10% of the SNPs. (The minimum of Fay and Wu's H lies around -30 for all plots.)