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

Alice Ledda[1], Guillaume Achaz[2,3,4], Thomas Wiehe[5] and Luca Ferretti[2,3,4,6]*

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

## 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 $\theta$ 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 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 measure of tree balance which is mostly determined by the root balance of the tree. We also compute the maximum and minimum values for neutrality tests as a function of sample size.

*Email: luca.ferretti@gmail.com

Focusing on the standard coalescent model of neutral evolution, we discuss how waiting times between coalescent events are related to derived allele frequencies and thereby to the frequency spectrum. Finally, we show how tree balance affects the frequency spectrum. In particular, we derive the complete SFS conditioned on the root imbalance. We show that the conditional spectrum is peaked at frequencies corresponding to the root imbalance and strongly biased towards rare alleles.

# 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 $\gamma$ statistic relates to the waiting times (Pybus *et al.*, 2000) and the $\beta$ 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 site frequency spectrum $\boldsymbol{\xi} = (\xi_1, ..., \xi_{n-1})$ of a sample of $n$ sequences is defined as the vector of counts $\xi_i$, $i \in \{1, ..., n-1\}$, 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. In absence of recombination, the SFS carries information on the realized coalescent tree and can be used to estimate tree structure (both waiting times and topology).

Recombination leads to a fragmentation of the sequences into haplotype blocks. Each such block has its own - albeit not independent - genealogy. The genealogies of neighboring blocks are strongly correlated, and sometimes even identical, depending on the type of the recombination event and its location in the tree. It can be shown that, under neutrality, most recombination events affect the lower part of the tree, while only a few affect upper branches with a "drastic" effect on the tree structure (Ferretti *et al.*, 2013).

A convenient model for the coalescent with a moderate number of recombination events is the ancestral recombination graph (ARG) (Griffiths and Marjoram, 1996; Wiuf and Hein, 1999), that tracks the genealogy of all haplotype blocks. At a genome scale, the number of recombination events is however enormous and, consequently, the number of haplotype blocks and associated genealogies is also extremely large. The SFS observed in simulated or experimental genome-scale data is therefore an average over many quasi-independent realizations of possible trees. There may be no single tree representative for the history of the entire sequences. Still, the SFS characterizes the 'average' coalescent tree, carrying information about average waiting times and topologies.

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). 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 coalescent trees upon the SFS. First, we will introduce the theoretical framework for neutrality tests and tree balance. Then, we will present the decomposition of the SFS in terms of waiting times and tree shape. We will start by 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 will then study the case of infinite recombination (average topology), providing a framework for genome-wide analyses. As an application, we will present the interpretation of neutrality tests in terms of tree topology and we will derive their maximum and minimum values. Finally, we will explicitly compute the impact of both the waiting times and the tree balance on the average SFS.

A qualitative summary of the results about the interpretation of neutrality test is given in Table 3.

# Theoretical framework

## Estimators of $\theta$ and neutrality tests

A fundamental population genetic quantity is the scaled mutation rate $\theta = 2pN_e\mu$, where $p$ is the ploidy (typically $p = 1$ or 2), $\mu$ is the mutation rate per generation per sequence and $N_e$ is the effective size of the population. $\theta$ 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 $\theta$ is Watterson's $\hat{\theta}_S = S/a_n$, where $S$ is the number of observed polymorphic sites in a sequence sample of size $n$ ("segregating sites"), and $a_n = \sum_{i=1}^{n-1} 1/i$ is the $(n-1)$th harmonic number (Watterson, 1975).

More generally, it has been shown that many of the well-known $\theta$-estimators can be expressed as linear combinations of the components $\xi_i$ of the SFS (Tajima, 1983; Achaz, 2009; Ferretti *et al.*, 2010). For example, $\hat{\theta}_S = \sum_{i=1}^{n-1} \frac{1}{a_n}\xi_i$ or Tajima's $\theta_\pi = \sum_{i=1}^{n-1} \frac{2i(n-i)}{n(n-1)}\xi_i$. 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 $\theta$-estimators, hence as a linear combination of the $\xi_i$. For instance, the non-normalized Tajima's $D$ (Tajima, 1989) is $\hat{\theta}_\pi - \hat{\theta}_S$, 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 $\xi_i$ 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 $\xi_i$ with large $i$ (high frequency derived alleles), while $D$ assigns negative weight to $\xi_i$ with small and large $i$ (rare alleles).

## Coalescent trees

The coalescent trees considered here are binary trees with labelled histories. This means that all internal nodes have two descendant lineages and are uniquely ordered by time. Such coalescent 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.

The waiting times between subsequent coalescent events, i.e. the level heights, are denoted by $t_k$. Under neutrality and constant population size 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

$$\begin{aligned} E(h) &= 1 - 1/n \quad \text{and} \\ E(l) &= a_n \, . \end{aligned}$$

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.

## Tree topology

**Branch size and its moments in fixed trees.** Following Fu (1995), we define the *size* $\sigma_k$ of a branch from level $k$ as the number of leaves that descend from that branch. Any mutation on this branch is carried by $\sigma_k$ sequences from the present sample. We denote by $P(\sigma_k = i|T)$ the probability that a randomly chosen branch of level $k$ is of size $i$, given tree $T$. The complete set of distributions $P(\sigma_k = i|T)$ 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 $E(\sigma_k) = \sum_{i=1}^{n-k+1} iP(\sigma_k = i|T) = n/k$. 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, $\mathrm{Var}(\sigma_k)$, 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. $\mathrm{Var}(\sigma_k)$ also varies greatly from level to level: for example, the variance of the uppermost level

$\mathrm{Var}(\sigma_2) \in [0, n^2/4 - n + 1]$, whereas $\mathrm{Var}(\sigma_n) = 0$ for all trees because $\sigma_n = 1$ 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

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

Minimum variance at level $k$ is obtained when all lineages have either[1] $\lfloor n/k \rfloor$ or $\lfloor n/k \rfloor + 1$ descendants and it is always $\leq 1/4$:

$$\min_T \mathrm{Var}(\sigma_k) = (n/k - \lfloor n/k \rfloor) \cdot (\lfloor n/k \rfloor + 1 - n/k) . \tag{2}$$

We propose an informative statistics on tree balance based on $\mathrm{Var}(\sigma_k)$. Given a random point on the tree, we can compute the variance in its branch size given its level. Then, we average the variance across the whole length of the tree. Fixing a tree $T$, the average variance in branch size across all levels $k$ is

$$\overline{\mathrm{Var}(\sigma)} = \frac{\sum_{k=2}^{n} k\, t_k \mathrm{Var}(\sigma_k)}{\sum_{k=2}^{n} k\, t_k} = \frac{1}{l} \sum_{k=2}^{n} k\, t_k \mathrm{Var}(\sigma_k) . \tag{3}$$

This summary statistic contains the natural weights $k\, t_k$, that is the amount of branch lengths at level $k$. 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.

**Random trees.** We consider trees generated by the equal-rates Markov (ERM) or Yule model (Yule, 1925), i.e. at a splitting event all lineages have the same probability to split.

The probability $P(\sigma_k^* = i|n)$ that a branch of level $k$ has size $i$ in a random ERM tree of total size $n$ is given by Fu (1995)

$$P(\sigma_k^* = i|n) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \tag{4}$$

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

As noted above, the mean is

$$\mathrm{E}(\sigma_k^*) = \sum_{i=1}^{n-1} i \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} = n/k\,,$$

independent of topology. Furthermore, the variance is

$$\mathrm{Var}(\sigma_k^*) = \frac{n(k-1)(n-k)}{k^2(k+1)} = \frac{k-1}{k+1}\left(\frac{n}{k}-1\right)\frac{n}{k} \tag{5}$$

and the average variance across levels becomes

$$\overline{\mathrm{Var}(\sigma^*)} = \frac{1}{l}\sum_{k=2}^{n} k\, t_k \mathrm{Var}(\sigma_k^*)\,. \tag{6}$$

**Node size and balance.** The notion of size can be easily extended to internal nodes. The *size* of an internal node, denoted $|\nu_k|$ for the node $\nu_k$ $(k \in (1, n-1))$, is the size of its incoming (i.e. parental) branch. For instance, the size of $\nu_1$ – the root – is $|\nu_1| = n$; the size of the most recent internal node is $|\nu_{n-1}| = 2$. The size of a node can be further divided into the number of left and right descendants, $\lambda_k$ and $\rho_k$, respectively. Let $\omega_k = \min(\lambda_k, \rho_k)$. Then, we have $1 \leq \omega_k \leq \lfloor |\nu_k|/2 \rfloor$. Unbalanced nodes have $\omega_k = 1$, completely balanced nodes have $\omega_k = \lfloor |\nu_k|/2 \rfloor$. Under neutrality $\omega_k$ are (quasi-)uniform random variables and their standardized sum is approximately normal. This fact can be used to construct tests of neutrality (Li and Wiehe, 2013).

## Decomposition of the Site Frequency Spectrum (SFS)

Let the vector $\boldsymbol{\xi} = (\xi_1, ..., \xi_{n-1})$ denote the site frequency spectrum (SFS). For each component $\xi_i$ the product $i\,\xi_i$ is an unbiased estimator of $\theta = 2pN\mu$. Hence, given weights $\boldsymbol{w} = (w_1, ..., w_{n-1})$, the weighted linear combination

$$\hat{\theta}_{w_i} = \frac{1}{\sum w_i}\sum_{i=1}^{n-1} w_i\, i\, \xi_i \tag{7}$$

is also an unbiased estimator of $\theta$. For instance, Watterson's estimator $\hat{\theta}_S = S/a_n$ follows from setting $w_i = 1/i$ in eq (7); Tajima's estimator $\hat{\theta}_\pi$ (Tajima, 1983) is obtained by letting $w_i = (n - i)$. In fact, one can write all usual $\theta$ 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.

**Given topology, given waiting times.** In this section we discuss the dependence of the average spectrum $\boldsymbol{\xi}$ on tree topology.

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 $\mu$ per unit time. Fixing a tree with respect to shape and branch lengths, we can average over the mutation process and obtain for the mean frequency spectrum (Fu, 1995)

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

The probabilities $P(\sigma_k = i|T)$ represent the distribution of $\sigma_k$, the number of descendants of the branches of level $k$, therefore they depend only on the shape of the tree $T$ and not on waiting times.

Replacing $\xi_i$ by their mean according to eq (8), we obtain the general expression for the mean of SFS-based $\theta$-estimators

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

Interestingly, several common estimators can be written in terms of a general weight function of the form $w_i = \alpha i + \beta + \gamma/i$ with appropriate values of $\alpha, \beta, \gamma$. For instance, $\hat{\theta}_S$ has $\alpha = \beta = 0$ and $\gamma = 1$, while $\hat{\theta}_\pi$ has $\alpha = -1$, $\beta = n$ and $\gamma = 0$. With this special weight function, equation (9) becomes

$$\mathrm{E}_\mu(\hat{\theta}_{\boldsymbol{w}}|T) = \frac{\theta}{N_{\alpha,\beta,\gamma,n}} \sum_{i=1}^{n-1} \sum_{k=2}^{n} (\alpha i^2 + \beta i + \gamma)\, k\, t_k\, P(\sigma_k = i|T)\,. \tag{10}$$

with $N_{\alpha,\beta,\gamma,n} = \alpha\frac{n(n-1)}{2} + \beta(n-1) + \gamma a_n$. Using $\sum_{i=1}^{n-1} iP(\sigma_k = i|T) = \mathrm{E}(\sigma_k) = n/k$ and $\sum_{i=1}^{n-1} i^2 P(\sigma_k = i|T) = \mathrm{Var}(\sigma_k) + \mathrm{E}^2(\sigma_k)$ and exchanging the order of the sums, this becomes

$$\mathrm{E}_\mu(\hat{\theta}_{\boldsymbol{w}}|T) = \frac{\theta}{N_{\alpha,\beta,\gamma,n}} \left( \alpha \overline{\mathrm{Var}(\sigma)} l + \sum_{k=2}^{n} t_k \left( \alpha \frac{n^2}{k} + \beta n + \gamma k \right) \right) \tag{11}$$

**Average topology, given waiting times.** To dissect the effect of branch lengths and topology on the SFS, we now assume an arbitrary ERM-generated tree and study the impact of coalescent waiting times on the SFS. This is applicable to temporal changes in population size, since they affect the waiting times but not the average tree topology. This will help us to understand the contributions of the different levels of the genealogy to the SFS, to the estimators of $\theta$ and to the neutrality tests derived from them.

From basic coalescent theory we know that $\mathrm{E}(t_k) = 1/k(k-1)$ for constant population size. The general case of non-constant population size is treated in Griffiths and Tavaré (1998), and formulae for the mean and covariance of waiting times are derived in Zivkovic and Wiehe (2008).

Genealogies of unlinked loci can be treated as independent replicates of the same process. When moving across the genome, trees from different (unlinked) loci are then samples from the same distribution. Since fluctuations in population size do not affect the distribution of tree shapes, the SFS pooled from different loci represents an average over topologies.

For the average topology, equation (8) has the simple form

$$\mathrm{E}(\xi_i | t_2 \dots t_n) = \theta \sum_{k=2}^{n} k\, t_k\, P(\sigma_k^* = i | n)\,. \tag{12}$$

Furthermore,

$$\mathrm{E}_\mu(\hat{\theta}_{\boldsymbol{w}} | t_2 \dots t_n) = \frac{\theta}{\sum w_i} \sum_{i=1}^{n-1} \sum_{k=2}^{n} i\, w_i\, k\, t_k\, P(\sigma_k^* = i | n) \tag{13}$$

As the variance of $\sigma_k^*$ can be explicitly computed under the ERM model from equation (6), setting the weight to $w_i = \alpha i + \beta + \gamma / i$ the previous equation reduces to:

$$\mathrm{E}_\mu(\hat{\theta}_{\boldsymbol{w}} | t_2 \dots t_n) = \frac{\theta}{N_{\alpha,\beta,\gamma,n}} \sum_{k=2}^{n} t_k \left( \alpha \frac{n(2n-k+1)}{k+1} + \beta n + \gamma k \right) \tag{14}$$

# Applications

## Estimators of $\theta$ and related neutrality tests

The above results lead to a simple interpretation of the usual test statistics in terms of tree shape and balance. We consider first the case of a given tree topology and then the average case. The tests are summarised in Table 2 and their interpretation in Table 3.

### Interpretation for a given genealogy

In this section we consider a single locus with a fixed genealogy.

**Tajima's $D$** statistic is the most used neutrality test. It is proportional to the difference $\hat{\theta}_\pi - \hat{\theta}_S$.

We compute Watterson's estimator from eq (11) by setting $\alpha = \beta = 0$ and $\gamma = 1$, obtaining

$$\mathrm{E}_\mu(\hat{\theta}_S) = \theta \frac{l}{a_n} \, . \tag{15}$$

Note that $\mathrm{E}_\mu(\hat{\theta}_S)$ is proportional to the total length of the tree, divided by the mean length. As such, it is independent from the tree topology.

Similarly, letting $\alpha = -1$, $\beta = n$ and $\gamma = 0$, Tajima's $\theta_\pi$ is

$$\mathrm{E}_\mu(\hat{\theta}_\pi) = \theta \frac{2}{n(n-1)} \left[ -\overline{\mathrm{Var}(\sigma)} l + n^2 \sum_{k=2}^{n} t_k (1 - 1/k) \right] . \tag{16}$$

Note that $\mathrm{E}_\mu(\hat{\theta}_\pi)$ can be decomposed into two components: one that is a linear combination of tree lengths, independent from the topology, plus a component that contains the quantity $\overline{\mathrm{Var}(\sigma)} l$. The latter is strongly related to tree balance. In fact, once "normalised" by $l$, it represents the average imbalance along the tree as explained before.

As Tajima's $D$ is the difference between these two estimators – up to normalization depending only on $n$ and $\theta$ – it can be re-expressed as:

$$\mathrm{E}_\mu(D) \propto -\frac{2}{n(n-1)} \overline{\mathrm{Var}(\sigma)} l + \sum_{k=2}^{n} t_k \left( \frac{2n}{(n-1)} \left( 1 - \frac{1}{k} \right) - \frac{k}{a_n} \right) . \tag{17}$$

In qualitative terms,

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

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. 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.

**Fay and Wu's** $H$ test 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). To compute $\mathrm{E}_\mu(H)$, we substitute $\alpha = 1$, $\beta = \gamma = 0$ in eq (11) and obtain

$$\mathrm{E}_\mu(\hat{\theta}_H) = \theta \frac{2}{n(n-1)} \left( \overline{\mathrm{Var}(\sigma)}\ l + n^2 \sum_{k=2}^{n} \frac{t_k}{k} \right) . \tag{18}$$

Then, the $H$ test has the property

$$\mathrm{E}_\mu(H) \propto -\frac{4}{n(n-1)} \overline{\mathrm{Var}(\sigma)}\ l + \frac{2n}{n-1} \sum_{k=2}^{n} t_k (1 - 2/k) . \tag{19}$$

Its qualitative interpretation is

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

Like Tajima's $D$, $H$ contains the imbalance term with negative sign. However, it has another contribution that weights negatively the waiting times close to the root and positively the waiting times close to the leafs – which is opposite to Tajima's $D$. 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 $E$** is another test designed to detect selective sweeps. However, it is known to be less powerful than $H$ (Zeng *et al.*, 2006). The estimator $\hat{\theta}_L$ is defined by setting $\beta = 1$ and $\alpha = \gamma = 0$. The mean is

$$\mathrm{E}_\mu(\hat{\theta}_L) = \theta \frac{n}{n-1} h \tag{20}$$

and the $E$-test $\hat{\theta}_L - \hat{\theta}_S$ has the property

$$\mathrm{E}_\mu(E) \propto \frac{n}{n-1} h - \frac{l}{a_n} \, . \tag{21}$$

The qualitative interpretation is

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

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, the qualitative analysis shows that $E$ compares upper and lower branches, i.e. height and length of the tree. Hence, $E$ can be naturally interpreted as a test for star-likeness of a tree. In star-like trees the length is $n$ times the height.

**Fu and Li's $D_{FL}$** is one of several tests based on singletons. Its mean is

$$\mathrm{E}_\mu(D_{FL}) \propto l - \sum_{k=2}^{n} k t_k P_{n,k}(1|T) \tag{22}$$

This test has the qualitative interpretation

$$\boxed{D_{FL} \simeq \ + \text{length of internal branches} \ - \ \text{length of external branches}} \tag{23}$$

It measures the relative contribution of external branches to total tree length. In complete star-like trees total length and external branch length are identical.

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 (23) is not valid anymore.

## Interpretation for average topology, given waiting times

For large genomic sequences – and with limited population structure – recombination effectively leads to an averaging over tree shapes. It is therefore interesting to determine the test statistics for such a scenario.

In this section we will present the interpretation of the tests for an average topology. The results are linear in the waiting times, therefore it is trivial to average over times as well, by substituting $t_k$ with $\mathrm{E}(t_k)$.

**Tajima's $D$:** Averaging over tree shapes, $\hat{\theta}_S$ does not change, as it does not depend on tree shape. In contrast, for $\hat{\theta}_\pi$ we have

$$\mathrm{E}_\mu(\hat{\theta}_\pi) = \theta \frac{2}{(n-1)} \sum_{k=2}^{n} t_k \frac{(n+1)(k-1)}{k+1} \,. \tag{24}$$

Thus, $D$ becomes

$$\mathrm{E}_\mu(D) \propto \theta \left[ 2\frac{n+1}{n-1} \sum_{k=2}^{n} t_k \frac{(k-1)}{k+1} - \frac{l}{a_n} \right] \tag{25}$$

The interpretation of $D$ is different once topologies are averaged: it weights positively the old branches and negatively the young branches, so it measures starlikeness similar to Zeng's $E$, but with less weight on the root branches.

**Fay and Wu's $H$:** $\hat{\theta}_H$ depends on tree shape. In case of an average tree it becomes

$$\mathrm{E}_\mu(\hat{\theta}_H) = \theta \frac{2}{n-1} \sum_{k=2}^{n} t_k \frac{2n-k+1}{k+1} \,. \tag{26}$$

Thus, the interpretation of Fay and Wu's $H$ is also affected. Now, we find that

$$\mathrm{E}_\mu(H) \propto \theta \frac{2}{n-1} \sum_{k=2}^{n} \frac{t_k}{k+1} \left[ n(k-3) + 2(k-1) \right] \,, \tag{27}$$

i.e. the young branches are weighted positively, while most negative weight is attributed to the root branches ($k=2$).

**Zeng's $E$:** Neither $\hat{\theta}_S$ nor $\hat{\theta}_L$ depend on shape, so the interpretation of Zeng's $E$ is the same as for fixed trees.

**Fu and Li's $D_{FL}$:** Once averaged over shapes, $\xi_1$ becomes

$$\mathrm{E}_\mu(\hat{\theta}_{\xi_1}) = \frac{\theta}{n-1} \sum_{k=2}^{n} k(k-1) t_k \tag{28}$$

and $D_{FL}$ tends to weight positively the younger branches, as expected:

$$\mathrm{E}_\mu(D_{FL}) \propto \theta \left( \frac{1}{n-1} \sum_{k=2}^{n} k(k-1) t_k - \frac{l}{a_n} \right) \tag{29}$$

### Extreme values of neutrality tests

In this section we derive a simple approximation for the extreme positive and negative values of neutrality tests.

From the results of the previous sections, it is easy to obtain the average values of the tests $\mathrm{E}_\mu(\hat{\theta}_{\boldsymbol{w}}|T,S)$ conditioned both on the tree $T$ and on the number of segregating sites $S$, simply by substituting $\theta$ with $S/l$. Then we are able to compute the maximum and minimum value of $\mathrm{E}_\mu(\hat{\theta}_{\boldsymbol{w}}|T,S)$, which correspond to the maximum and minimum value of the statistics neglecting the Poisson mutational noise.

The numerators of the tests depend linearly on $\theta$ and therefore on $S$, while the denominators of the tests are of the form $\sqrt{\alpha_n S + \beta_n S(S-1)}$ (Tajima, 1989; Zeng *et al.*, 2006).

The maximum and minimum of the tests across all trees $T$ depend on $n$ and $S$; however, for large $S$, they depend only on the sample size $n$. The extreme values are presented in Figure 2 as a function of $n$ and for different values of $S$.

Since the tests are distributed with mean 0 and variance 1 under the null neutral model, a maximum close to or less than 2 would suggest that the distribution is compressed around its maximum (i.e. the upper bound of any confidence interval would fall close to the maximum itself) and the test would probably have reduced power on the positive tail. Analogous reasoning is valid for minimum values close to or larger than $-2$.

In the derivations of this section we neglect the contribution of minimum imbalance, $\min_T \overline{\mathrm{Var}(\sigma)}$, approximating it with 0.

**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 maximally unbalanced trees with length concentrated in the upmost and lowest branches ($k = 2, n$).

$$\max_T \mathrm{E}_\mu(D|T,S) = \frac{\left(\frac{n}{2(n-1)} - \frac{1}{a_n}\right) S}{\sqrt{\alpha_n^D S + \beta_n^D S(S-1)}} \xrightarrow[S\gg 1]{} \frac{\frac{n}{2(n-1)} - \frac{1}{a_n}}{\sqrt{\beta_n^D}} \xrightarrow[n\gg 1]{} \frac{3}{2\sqrt{2}} \log(n) \tag{30}$$

$$\min_T \mathrm{E}_\mu(D|T,S) = \frac{\left(\frac{2}{n} - \frac{1}{a_n}\right) S}{\sqrt{\alpha_n^D S + \beta_n^D S(S-1)}} \xrightarrow[S\gg 1]{} \frac{\frac{2}{n} - \frac{1}{a_n}}{\sqrt{\beta_n^D}} \xrightarrow[n\gg 1]{} -\frac{3}{\sqrt{2}} \approx -2.1 \tag{31}$$

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.

Our results suggest that Tajima's $D$ could be a very good test to detect balanced trees with long upper branches, but is not as good in detecting unbalanced trees with long upmost and/or lowest branches. It is actually well-known that its null distribution is skewed and compressed for negative values. However, our results suggest that it is generally difficult to detect significant deviations from the Kingman coalescent by negative values of Tajima's $D$, unless there is an excess of rare mutations due to non-homogeneous processes (e.g. background/purifying selection).

**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 maximally unbalanced trees with length concentrated in the upmost branches ($k = 2$).

$$\max_T \mathrm{E}_\mu(H|T,S) = \frac{\frac{n}{4(n-1)}S}{\sqrt{\alpha_n^H S + \beta_n^H S(S-1)}} \xrightarrow[S\gg1]{} \frac{\frac{n}{4(n-1)}}{\sqrt{\beta_n^H}} \xrightarrow[n\gg1]{} \frac{\log(n)}{4\sqrt{\pi^2 - 88/9}} \tag{32}$$

$$\min_T \mathrm{E}_\mu(H|T,S) = \frac{-\frac{(n-2)^2}{n(n-1)}S}{\sqrt{\alpha_n^H S + \beta_n^H S(S-1)}} \xrightarrow[S\gg1]{} -\frac{\frac{(n-2)^2}{n(n-1)}}{\sqrt{\beta_n^H}} \xrightarrow[n\gg1]{} -\frac{\log(n)}{\sqrt{\pi^2 - 88/9}} \tag{33}$$

Our results suggest that Fay and Wu's $H$ test works very well at both extremes, especially for maximally unbalanced trees.

**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 $k = n$).

$$\max_T \mathrm{E}_\mu(E|T,S) = \frac{\left(\frac{n}{2(n-1)} - \frac{1}{a_n}\right)S}{\sqrt{\alpha_n^E S + \beta_n^E S(S-1)}} \xrightarrow[S\gg1]{} \frac{\frac{n}{2(n-1)} - \frac{1}{a_n}}{\sqrt{\beta_n^E}} \xrightarrow[n\gg1]{} \sqrt{\frac{3}{\pi^2 - 9}}\log(n) \tag{34}$$

$$\min_T \mathrm{E}_\mu(E|T,S) = \frac{\left(\frac{1}{n-1} - \frac{1}{a_n}\right)S}{\sqrt{\alpha_n^E S + \beta_n^E S(S-1)}} \xrightarrow[S\gg1]{} \frac{\frac{1}{n-1} - \frac{1}{a_n}}{\sqrt{\beta_n^E}} \xrightarrow[n\gg1]{} -\sqrt{\frac{3}{\pi^2 - 9}} \approx -1.9 \tag{35}$$

Our results suggest that Zeng's $E$ test could work well to detect tree with long upper branches, but not so well to detect star-like trees.

## Impact of the waiting times on the SFS

The decomposition of the SFS can also be interpreted in terms of the frequency spectrum itself, instead of neutrality tests. When averaging over tree shapes, the frequency spectrum is directly informative with respect to the waiting times $t_k$.

For an average topology, the relation between the mean spectrum and waiting times is given by eq (12). The average variation in the spectrum due to a small change $\delta t_k$ in time is

$$\mathrm{E}\left[\frac{\delta \xi_i}{\delta t_k}\right] = \theta k \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} . \tag{36}$$

The variation due to a *relative* change in $t_k$ is

$$\mathrm{E}\left[\frac{\delta \xi_i}{\delta t_k / t_k}\right] = \theta k \mathrm{E}(t_k) \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \tag{37}$$

This expression represents the sensitivity of the spectrum to a relative change in waiting times. Interestingly, it also corresponds to the average contribution of the mutations at level $k$ to the $i$th component of the frequency spectrum. For constant population size, we have

$$\mathrm{E}\left[\frac{\delta \xi_i}{\delta t_k / t_k}\right] = \theta \frac{\binom{n-i-1}{k-2}}{(k-1)\binom{n-1}{k-1}} \tag{38}$$

This captures the sensitivity of the spectrum to a relative change in waiting times with respect to the case of constant population size.

The relative sensitivity (equation 38) is shown in Figure 1 as a function of frequency $i$ and level $k$. Obviously, information about the waiting times for the lowest levels, close to the leaves, are mostly contained in the number of rare mutations. In other words, a change in the waiting times of lower parts (e.g. $t_n$ or $t_{n-1}$) would mostly impact the low frequency components of the SFS. These components of the spectrum are also sensitive to relative variations of all other waiting times. On the contrary, variations on the times close to the root is spread across all frequencies, but is dominant for high frequencies. A change in the waiting time $t_2$ would impact all frequencies of the SFS equally – but a change in the highest component of the frequency spectrum can be unequivocally traced to a variation of that waiting time.

## Impact of the topology on SFS

To better characterize the impact of topology on SFS, we first decomposed $\overline{\mathrm{Var}(\sigma)}$ into the contributions from different levels. These contributions can be considered either as contributions per unit time – i.e. variances are weighted by the number of lineages $k$ – or contributions per level – i.e. variances are weighted by the length at level $k$, $k\mathrm{E}(t_k)$, which is $1/(k-1)$ for constant population size. We show the contributions in Figure 3. It is clear that the largest contributions to $\overline{\mathrm{Var}(\sigma)}$ come from the levels close to the root. In particular, the dominant contribution from the uppermost level depends strongly on the root balance $\omega_1$, which has been previously recognised as a meaningful global measure of tree balance (Ferretti *et al.*, 2013; Li and Wiehe, 2013).

Since $\overline{\mathrm{Var}(\sigma)}$ is dominated by the root balance $\omega_1$, we consider now the mean frequency spectrum conditioned on $\omega_1$. The equation for the mean SFS, averaged over waiting times, shape and the mutation process, but conditioned on $\omega_1$, is

$$\mathrm{E}(\xi_i|\omega_1) = \frac{\sum_{\{T|\Omega_1(T)=\omega_1\}} \mathrm{E}(\xi_i|T)P(T)}{\sum_{\{T|\Omega_1(T)=\omega_1\}} P(T)} = \theta \sum_{k=2}^{n} k\mathrm{E}(t_k) \frac{\sum_{\{T|\Omega_1(T)=\omega_1\}} P(\sigma_k = i|T)P(T)}{\sum_{\{T|\Omega_1(T)=\omega_1\}} P(T)}$$
$$= \theta \sum_{k=2}^{n} k\mathrm{E}(t_k) P(\sigma_k = i|\omega_1)\,. \tag{39}$$

The distribution of the number of descendants $P(\sigma_k = i|\omega_1)$ has been obtained in Ferretti *et al.* (2013). It is based on combinations of probabilities for the number of descendants of a set of lineages, using the theory of Polya urn models. These probabilities are then summed over the unknown number of left $(x)$ and right $(k-x)$ descendants of the root at level $k$:

$$P(\sigma_k = i|\omega_1) = \sum_{x=1}^{\omega_1} \left( P(i|x,\omega_1)\frac{x}{k} + P(i|k-x, n-\omega_1)\frac{k-x}{k} \right) P(x|\omega_1,k,n)\,, \tag{40}$$

where

$$P(i|x,\omega_1) = \frac{\binom{\omega_1-i-1}{x-2}}{\binom{\omega_1-1}{x-1}} + \delta_{i,\omega_1}\delta_{x,1} \tag{41}$$

and $P(x|\omega_1,k,n) = \mathrm{Hyp}_{\omega_1-1,k-2;n-2}(x-1)$ is the hypergeometric distribution. (An alternative, closed form of (40) can be found in Ferretti *et al.* (2013)).

The resulting spectrum for constant population size is plotted in Figure 4 for different values of $\omega_1$. For any fixed value of $\omega_1$, the spectrum has two strong peaks at $i = \omega_1$ and $i = n - \omega_1$. The rest of the spectrum is dominated by rare alleles at low frequencies $i < \omega_1$. Curiously, in this range of frequencies, it shows a universal behaviour (independent on $\omega_1$). For $\omega_1 < i < n - \omega_1$, the spectrum is slightly more biased towards rare alleles than the neutral unconditioned spectrum $\theta/i$, while there are no mutations with $i > n - \omega_1$.

# 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 multilocus 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 nonrecombining 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 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 the SFS summary statistics and the 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 $\theta_S$, 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 random mutation – plays an important role in the interpretation of these neutrality tests. The simplicity of these results stem 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 3. Our results are rigorous and consistent with intuition. Our methods help to understand the peculiarities of the different tests. For example, we re-interpret Zeng's $Z$ 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 $\overline{\mathrm{Var}(\sigma)}$ is also related to other balance statistics proposed recently, namely the root balance $\omega_1$ 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.

We have also characterized the sensitivity of the SFS to waiting times at different levels and how the spectrum changes with root imbalance. For the average topology, we show that singletons and other rare mutations can be born at any level, while high frequency mutations occurred most probably close to the root and are therefore sensitive to the first waiting times only.

On the other hand, we have shown that the spectrum of highly unbalanced trees tends to be biased towards rare mutations, with a strong excess of mutations at frequencies corresponding to the numbers of left and right descendants of the root. This excess of mutations corresponds exactly to mutations located on the root branches of the genealogical tree. While these results were obtained conditioning the coalescent on the root balance $\omega_1$, it would be interesting to find an explicit algorithm to build coalescent trees with a given (root) balance.

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. Existing tests are sensitive to the variances $\mathrm{Var}(\sigma_k)$, but one could imagine other tests sensitive e.g. to the skewness or kurtosis of $P(\sigma_k = i|T)$ or other combinations. While the variance is a direct measure of imbalance and especially to the imbalance of the upper branches, other combinations could be sensitive to different features of the tree. Our results on extreme values could also give some indication about the effectiveness of these combinations. Alternatively, our approach can be used to understand the average structure of the genealogical trees generated by models for which the expected SFS is known.

Finally, some of our results could 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. 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 acknowl-

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

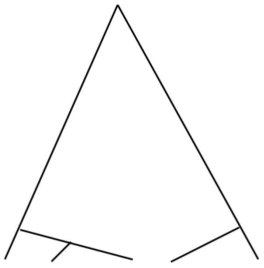
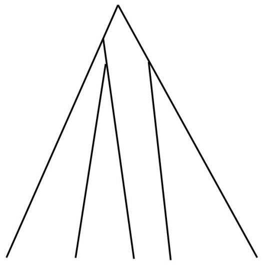
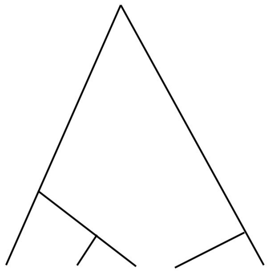
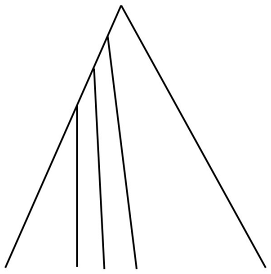
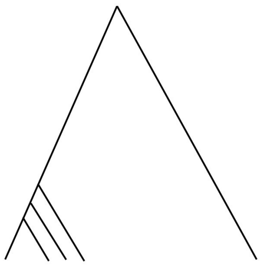
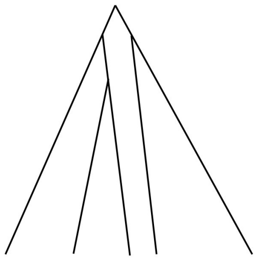
Table 1: Selected unbiased linear estimators of $\theta$.

| Test | formula | weights $w_i$ | reference |
|---|---|---|---|
| $D$ | $\hat{\theta}_\pi - \hat{\theta}_S$ | $i(n-i)/\binom{n}{2} - 1/a_n$ | Tajima (1989) |
| $H$ | $\hat{\theta}_\pi - \hat{\theta}_H$ | $i(n-2i)/\binom{n}{2}$ | Fay and Wu (2000) |
| $E$ | $\hat{\theta}_L - \hat{\theta}_S$ | $i/(n-1) - 1/a_n$ | Zeng *et al.* (2006) |
| $D_{FL}$ | $\hat{\theta}_{\xi_1} - \hat{\theta}_S$ | $\delta_{i,1} - 1/a_n$ | Fu and Li (1993) |

Table 2: Neutrality tests discussed in this paper.

edges support from DFG-SPP1590 by the German Science Foundation.

Table 3: Interpreting neutrality tests

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

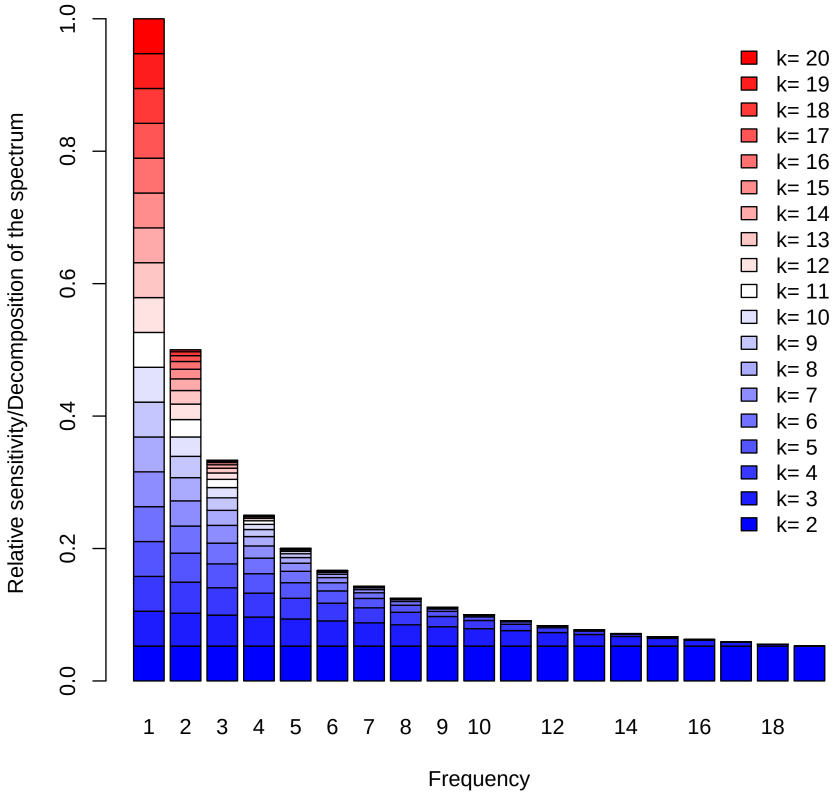


Figure 1: Barplot of the relative sensitivity $\delta\xi_i/(\delta t_k/t_k)$ of the different components of the spectrum to the waiting times of different levels $k = 2\ldots20$, for a sample with $n = 20$ and $\theta = 1$. This corresponds also to the decomposition of the spectrum in contributions from the different levels $k$.

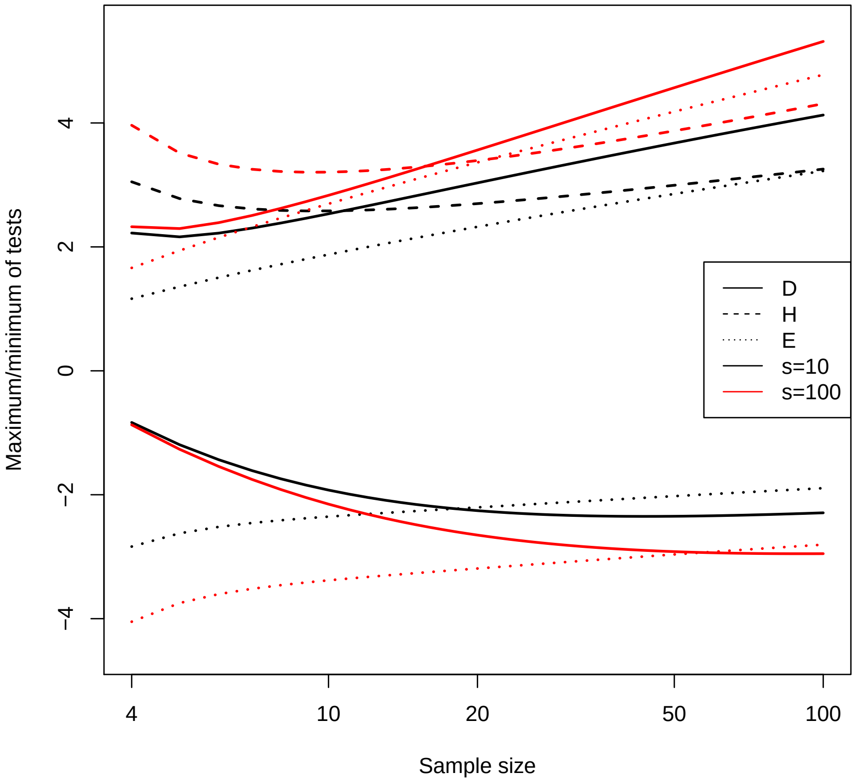


Figure 2: Maximum and minimum values of neutrality tests as a function of $n$ for $S = 10, 100$. The minimum of Fay and Wu's $H$ is not shown since its decreases from about $-10$ to $-30$ in the range of sample sizes of the plot.

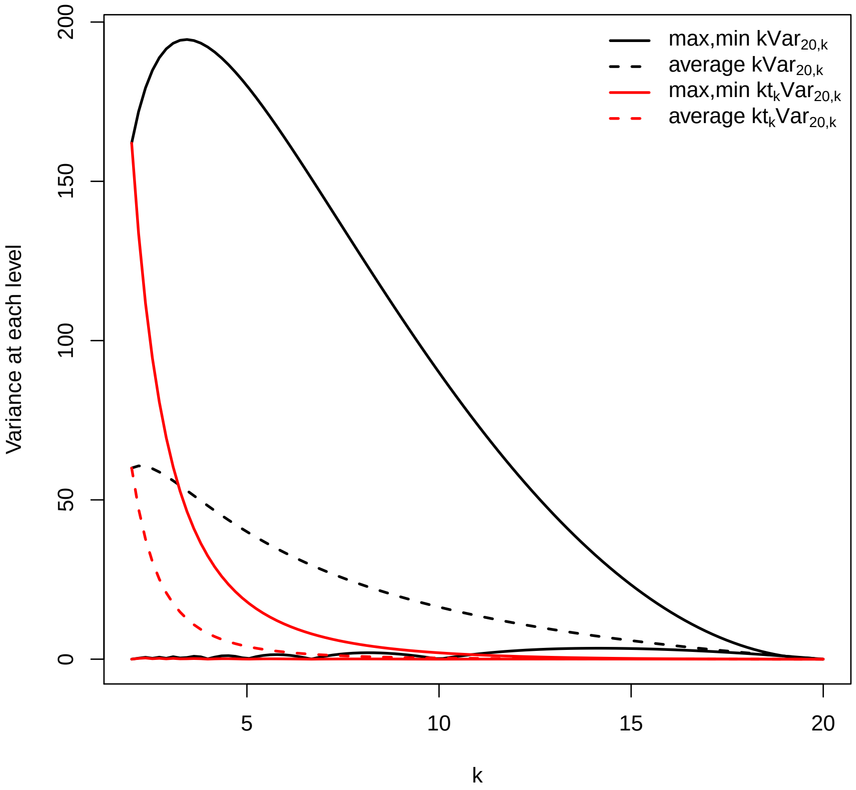


Figure 3: Plot of the mean, maximum and minimum contributions of different levels $k = 2 \ldots 20$ to the variance $\overline{\text{Var}(\sigma)} \cdot l$, for a sample with $n = 20$. In black the contribution per unit waiting time; in red, the total contribution per level.

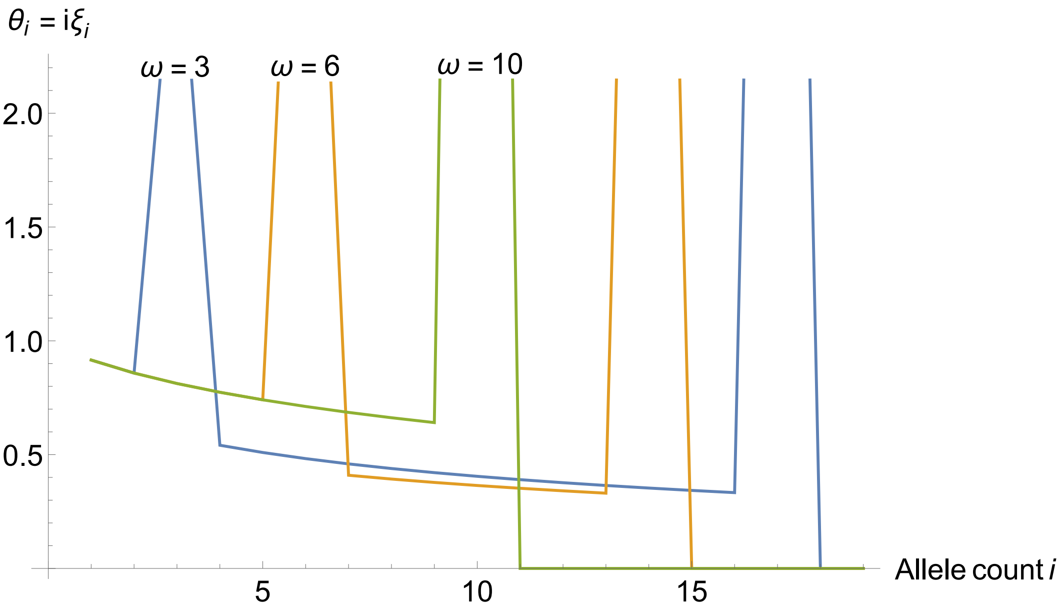


Figure 4: Normalized frequency spectrum $\xi_i(\omega)/(\theta/i)$ as a function of $i$ for $\omega_1 = 3, 6, 10$ and $n = 20$.

# Literature Cited

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

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.

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.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of dna sequences with recombination. Journal of Computational Biology **3**: 479–502.

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.

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.

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 &amp; Company Publishers Greenwood Village, Colorado.

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

Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theoretical population biology **55**: 248–259.

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.

# A Derivation of the variance of $\sigma_k^*$

## Some properties of the Binomial coefficients

The Upper Summation of the Binomial coefficients states that:

$$\sum_{i=k}^{n} \binom{i}{k} = \binom{n+1}{k+1} \tag{42}$$

From there, one can easily show that:

$$\sum_{i=k}^{n} \binom{i}{k} = \frac{1}{k} \sum_{i=k}^{n} i \binom{i-1}{k-1} \tag{43}$$

and therefore that:

$$\sum_{i=k}^{n} i\binom{i-1}{k-1} = k\binom{n+1}{k+1} \tag{44}$$

$$\sum_{i=k}^{n}(i+1)\binom{i}{k} = (k+1)\binom{n+2}{k+2} = \frac{(n+2)(k+1)}{(k+2)}\binom{n+1}{k+1} \tag{45}$$

$$\sum_{i=k}^{n} i\binom{i}{k} = \left(\frac{(n+2)(k+1)}{(k+2)} - 1\right)\binom{n+1}{k+1} \tag{46}$$

With a similar logic, one can show that:

$$\sum_{i=k}^{n} i^2\binom{i-1}{k-1} = k\sum_{i=k}^{n} i\binom{i}{k} \tag{47}$$

$$\sum_{i=k}^{n}(i+1)^2\binom{i}{k} = \left\{\frac{(n+2)(k+1)}{(k+2)}\left[\frac{(n+3)(k+2)}{(k+3)} - 1\right]\right\}\binom{n+1}{k+1} \tag{48}$$

## Mean and variance of $\sigma_k^*$

We want to derive the mean and variance of $\sigma_k^*$, which distribution is given by:

$$P(\sigma_k^* = i|n) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \tag{49}$$

Setting $X = n - i - 1$ and using the above equations, the mean becomes:

$$\begin{aligned}
E(\sigma_k^*) &= \frac{1}{\binom{n-1}{k-1}} \sum_{i=1}^{n-k+1} i\binom{n-i-1}{k-2} \\
&= \frac{1}{\binom{n-1}{k-1}} \sum_{X=k-2}^{n-2} [n-(X+1)]\binom{X}{k-2} \\
&= \frac{1}{\binom{n-1}{k-1}} \left[n\sum_{X=k-2}^{n-2}\binom{X}{k-2} - \sum_{X=k-2}^{n-2}(X+1)\binom{X}{k-2}\right] \\
&= n + \frac{n(k-1)}{k} \qquad (50) \\
&= \frac{n}{k} \qquad (51)
\end{aligned}$$

Similarly, the second moment of the distribution is:

$$\begin{aligned}
E(\sigma_k^{*2}) &= \frac{1}{\binom{n-1}{k-1}} \sum_{i=1}^{n-k+1} i^2 \binom{n-i-1}{k-2} \\
&= \frac{1}{\binom{n-1}{k-1}} \sum_{X=k-2}^{n-2} \left[n-(X+1)\right]^2 \binom{X}{k-2} \\
&= \frac{1}{\binom{n-1}{k-1}} \left[n^2 \sum_{X=k-2}^{n-2} \binom{X}{k-2} - 2n \sum_{X=k-2}^{n-2} (X+1)\binom{X}{k-2} + \sum_{X=k-2}^{n-2} (X+1)^2 \binom{X}{k-2}\right] \\
&= n^2 - 2n\frac{n(k-1)}{k} + \frac{n(k-1)}{k}\left[\frac{(n+1)k}{(k+1)} - 1\right] \\
&= \frac{n}{k(k+1)}(2n-k+1) \qquad (52)
\end{aligned}$$

It follows that:

$$\begin{aligned}
\mathrm{Var}(\sigma_k^{*2}) &= \frac{n}{k(k+1)}(2n-k+1) - \frac{n^2}{k^2} \qquad (53) \\
&= \frac{n(k-1)(n-k)}{k^2(k+1)} \qquad (54)
\end{aligned}$$