# The expected neutral frequency spectrum of linked sites

Luca Ferretti[1,2,3]*, Alexander Klassmann[4], Emanuele Raineri[5], Thomas Wiehe[4], Sebastian E. Ramos-Onsins[6], Guillaume Achaz[2,3]

(1) The Pirbright Institute, United Kingdom. (2) Atelier de Bioinformatique, ISyEB (UMR 7205 CNRS-MNHN-UPMC-EPHE), Paris, France (3) Stochastic Models for the Inference of Life Evolution, CIRB (UMR 7241 CNRS-INSERM), Collège de France, Paris. (3) (4) Institut für Genetik, Universität zu Köln, 50674 Köln, Germany. (5) CNAG-CRG, Centre for Genomic Regulation (CRG) and Universitat Pompeu Fabra (UPF), Barcelona, Spain. (6) Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain.

## Abstract

We present an exact, closed expression for the expected neutral Site Frequency Spectrum for two neutral sites, 2-SFS, without recombination. This spectrum is the immediate extension of the well known single site $\theta/f$ neutral SFS. Similar formulae are also provided for the case of the expected SFS of sites that are linked to a focal neutral mutation of known frequency. Formulae for finite samples are obtained by coalescent methods and remarkably simple expressions are derived for the SFS of a large population, which are also solutions of the multi-allelic Kolmogorov equations. Besides the general interest of these new spectra, they relate to interesting biological cases such as structural variants and introgressions. As an example, we present the expected neutral frequency spectrum of regions with a chromosomal inversion.

*Email: luca.ferretti@gmail.com

# 1 Introduction

One of the major features that characterizes nucleotide polymorphisms is the Site Frequency Spectrum (SFS), that is the distribution of the mutation frequencies at each site. The SFS can be computed either for the whole (large) population, assuming that the frequency $f$ is a continuous value in $(0, 1)$ or for a sample of $n$ individuals, for which the frequency is a discrete variable $f = k/n$, where $k \in [1, n-1]$. Typically sites with alleles at frequency 0 or 1 are not included in the SFS.

According to the standard neutral model of molecular evolution (Kimura, 1983), polymorphisms segregating in a population eventually reach a mutation-drift equilibrium. In this model, the expected neutral spectrum is proportional to the inverse of the frequency (Wright, 1938; Ewens, 2012). Using coalescent theory, Fu (1995) derived the mean and covariance matrix for each bin of the sample SFS, by averaging coalescent tree realizations across the whole tree space. For a single realization of the coalescent tree, results are different and depend on the realization; for example, mutations of high frequencies can be present only for highly unbalanced genealogies (Ledda *et al.*, 2015). The SFS was also studied in scenarios including selection (Fay and Wu, 2000; Kim and Stephan, 2002), demography (Griffiths and Tavaré, 1994; Živković and Wiehe, 2008) or population structure (Alcala *et al.*, 2016).

Besides its general interest, the SFS has been used to devise goodness-of-fit statistical tests to estimate the relevance of the standard neutral model for an observed dataset. SFS-based neutrality tests contrast estimations of the nucleotide variability from different bins of the sample SFS (Tajima, 1989; Fu and Li, 1993; Achaz, 2009). Ferretti *et al.* (2010) showed that, once the SFS under an alternative scenario (e.g. selection, demography or structure) is known, the optimal test to reject the standard neutral model is based on the difference between the standard neutral SFS and the alternative scenario SFS. All these tests assume complete linkage among variants in their null model.

Assuming independence between the sites, the observed SFS can also be used to estimate model parameters. An interesting recent approach is the estimation of piece-wise constant demography from genomewide SFS (e.g. Liu and Fu (2015)). More sophisticated methods based on the expected SFS, such as Poisson Random Field (Sawyer and Hartl, 1992; Bustamante *et al.*, 2001, 2002) and Composite Likelihood approaches (*e.g.,* Kim and Stephan, 2002; Li and Stephan, 2005; Kim and Nielsen, 2004; Nielsen *et al.*, 2005), have also played an important role in the detection of events of selection across regions of the genome. However, the assumption of linkage equilibrium is often violated in genetic data. In fact, while the average spectrum is insensitive to recombination, the presence of linked variants affects the distribution of summary statistics, therefore the spread (and possibly the mean) of the estimated parameters (Hudson *et al.*, 1990; Thornton, 2005). For this reason, simulations of the evolution of linked sequences are required for an accurate estimation of the statistical support for different models (Gutenkunst *et al.*, 2009).

The joint SFS for multiple sites has been the subject of longstanding investigations. The simplest spectrum for multiple sites is the "two-locus frequency spectrum" (Hudson, 2001), which we name the "two-Sites Frequency Spectrum" or 2-SFS. Assuming independence between the sites (*i.e.* free recombination), it simply reduces to the random association between two single-sites spectra (1-SFS). For intermediate recombination, a recursion solvable for small sample size has been provided (Golding, 1984; Ethier and Griffiths, 1990) as a well as a numerical solution relying on simulations (Hudson, 2001). Without recombination, finding an analytical expression for the spectrum has proven to be difficult.

There is a close relation between the $m$-SFS (the joint SFS of $m$ sites) and the multi-allelic spectrum of a single *locus* (defined as a sequence with one or more sites). Under the *infinite-sites* model, sites are assumed to have at most two alleles as new mutations occur exclusively at non-polymorphic sites. At the locus scale, each haplotype (the specific combination of the alleles carried at each point) can be interpreted as a single allele at a multi-allelic locus. In the absence of recombination, each point mutation either leaves the number of different haplotypes unchanged or generates one new haplotype. Therefore, at least conceptually, the SFS for $m$ biallelic sites at low mutation rate is closely related to the spectrum of $m+1$ alleles in a multi-allelic locus. Indeed, it is possible to retrieve the latter from the former by considering the $m+1$ alleles that result from the $m$ polymorphic sites. However, the $m$-SFS contains extra-information on the different linkage between the sites that is not available in the multi-allelic locus spectrum.

For an infinite population, the multi-alleles single-locus spectrum is the solution of a multiallelic diffusion equation (Ewens, 2012). Polynomial expansions were proposed to solve the diffusion equations for the SFS of an infinite population (Kimura, 1956; Littler and Fackerell, 1975; Griffiths, 1979). Finally, a polynomial expansion of the 2-SFS has been found for two sites without recombination and with general selection coefficients (Xie, 2011). However, the reported solution is an infinite series and is in sharp contrast with the simplicity of the solution for a single neutral site: $\mathrm{E}[\xi(f)] = \theta/f$. Furthermore, no closed form was provided for the 2-SFS of a sample.

Using a coalescent framework, the probability and size of two nested mutations were expressed by Hobolth and Wiuf (2009) as sums of binomial coefficients. Their formulae can be rewritten as an expected SFS in terms of a finite series. However their conditioning on exactly two nested mutations skews the spectrum and simulations show that even under this condition their result is valid only for $L\theta \ll 1$. Interesting analytical results on the spectrum of tri-allelic loci and recurrent mutations were obtained by Song and collaborators (Jenkins and Song, 2011; Jenkins *et al.*, 2014) for the Kingman coalescent and general allelic transition matrices. More recently, Sargsyan (2015) generalized the result of Hobolth and Wiuf (2009) by conditioning on any two mutations (nested or not) and extending it to populations of variable size. Moreover, he clarified the notion and classification of the 2-SFS.

In this work, we present a simple closed-form solution for the expectation of the neutral 2-SFS without recombination, for both the discrete sample 2-SFS and the continuous population 2-SFS.

The solution for a finite sample is obtained in a coalescent framework (Fu, 1995; Ferretti *et al.*, 2012) and its extrapolation to the limit of infinite sample sizes yields the continuous spectrum. Furthermore, we derive the expected 1-SFS of sites that are completely linked to a focal mutation of known frequency. In the appendices we also extend our results on the 2-SFS into closed expressions for the multi-allelic spectrum of a locus with three alleles.

Finally, as an application, we present exact results for the expected spectrum of neutral, non-recombining inversions. Chromosomal inversions are structural variants that play an important role in the adaptive evolution of some species (Hoffmann *et al.*, 2004), the most well-known case being flies in the *Drosophila* genus (Krimbas and Powell, 1992; Corbett-Detig and Hartl, 2012). We derive the expected frequency spectrum of neutral mutations linked to a neutrally evolving chromosomal inversion or a structural variant with similar properties. The neutral spectrum of inversions is more complex than the usual site frequency spectrum and represents the null model to detect population genetics signatures of selection on chromosomal variants (Kennington *et al.*, 2006; White *et al.*, 2009).

## Model definition and notation

We consider a population of size $N$ of haploid individuals without recombination. All subsequent results can be applied to diploids, provided that $2N$ is used instead of $N$, and to other cases by substituting the appropriate effective population size. We denote by $\mu$ the mutation rate per site and by $\theta = 2N\mu$ the population-scaled mutation rate per site. We work in the infinite-sites approximation, that is valid in the limit of small mutation rate $\theta \ll 1$. More properly, our results are derived in the limit $\theta \to 0$ with fixed non-zero $\theta L$, where $L$ is the length of the sequence. The expected value E[.] denotes the expectation with respect to the realizations of the evolutionary process for the sequences in the sample or in the whole population. We use *mutation* as a synonym for derived allele.

### Connection between sample and population SFS

We denote by $\xi(f)$ the *density* of mutations at frequency $f$ in the whole population and by $\xi_k$ the *number* of mutations at frequency $k/n$ in a sample of size $n$. Importantly, in both cases $f$ or $k$ refer to the frequency of the mutation, *i.e.* of the *derived* allele, and thus $\xi$ corresponds to the *unfolded* SFS.

The two spectra (sample and population) are related. Assuming that a mutation has frequency $f$ in the population, the probability of having $k$ mutant alleles in a random sample of size $n$ is simply given by the Binomial $\binom{n}{k} f^k (1-f)^{n-k}$. As the expected density of mutations at frequency $f$ in the population is given by $\mathrm{E}[\xi(f)]$, one can easily derive the sample frequency from the population frequency using the following sampling formula:

$$\mathrm{E}[\xi_k] = \int_{\frac{1}{N}}^{1-\frac{1}{N}} \binom{n}{k} f^k (1-f)^{n-k}\, \mathrm{E}[\xi(f)]\, df \qquad (1)$$

assuming that $n \ll N$.

Conversely, the population SFS can be derived from the sample SFS using the limit of large sample size $n \to \infty$. For a sample of $n$ individuals, the interval between the frequency bins is $1/n$ and therefore the density of mutations at the continuous frequency $f = k/n$ can be approximated[1] by $\mathrm{E}\left[\xi\left(\frac{k}{n}\right)\right] \approx \frac{\mathrm{E}[\xi_k]}{1/n} = n\mathrm{E}[\xi_k]$. The expected population spectrum can then be constructed from the limit:

$$\mathrm{E}[\xi(f)] = \lim_{n\to\infty} n\mathrm{E}[\xi_{\lfloor nf \rfloor}] \qquad (2)$$

for frequencies not too close to $\frac{1}{N}$ or $1 - \frac{1}{N}$.

[1] More formally, eq.(2) can be obtained from eq.(1) under the assumptions that $\frac{1}{N} \ll f, 1-f$ and that the population SFS is smooth over a range of frequencies $\Delta f \sim \frac{1}{N}$.

For a sample of size $n$, the expected neutral spectrum for constant population size is $\mathrm{E}[\xi_k] = \theta L/k$ and consequently, we have $\mathrm{E}[\xi(f)] = \theta L/f$ (Wright, 1938; Ewens, 2012). These results are exact for the Kingman coalescent and the diffusion equations respectively, and they are approximately valid for neutral models for frequencies $f \gg \frac{1}{N}$. For frequencies of order $\frac{1}{N}$, model-dependent corrections are needed and equation (2) is not valid anymore.

In the rest of this section we will deal with sample and population spectra together. We will slightly abuse the notation and switch between number and density of mutations, or probability and probability density.

## Conditional 1-SFS and joint 2-SFS

In the following, we will use two related but different kinds of spectra.

The first one is the joint 2-SFS of two bi-allelic sites. It is denoted $\xi(f_1, f_2)$ for the population and $\xi_{k,l}$ for the sample. It is defined as the density of pairs of sites with mutation frequencies at $f_1$ and $f_2$ for the population (resp. $k/n$ and $l/n$ for the sample). This is a natural generalization of the classical SFS for a single site. The expected spectrum $\mathrm{E}[\xi(f_1, f_2)]$ has two equivalent interpretations in the small $\theta$ limit: (a) for a sequence, it is the expected density of pairs of sites that harbor mutations with frequencies $f_1$ and $f_2$; (b) for two randomly chosen linked polymorphic sites, it is the probability density that they contain mutations with frequencies $f_1$ and $f_2$.

The second one is a conditional 1-SFS, a frequency spectrum of sites that are linked to a focal mutation of frequency $f_0$. It is denoted $\xi(f|f_0)$ for the population and $\xi_{k|l}$ for the sample. Again, this spectrum represents both (a) the expected density of single-site mutations of frequency $f$ in a locus linked to a focal neutral mutation of frequency $f_0$ and (b) the probability density that a randomly chosen site (linked to the focal site) hosts a mutation at frequency $f$.

Note that despite the similarity in notation, the two spectra $\xi(f, f_0)$ and $\xi(f|f_0)$ are different. The difference is the same as the one between the *joint probability* $p(f, f_0)$ that two sites $x$ and $x_0$ have mutations of frequency $f$

and $f_0$ respectively, and the *conditional probability* $p(f|f_0)$ that a mutation at site $x$ has frequency $f$ given that there is a mutation of frequency $f_0$ at a focal linked site $x_0$. Furthermore, the joint spectrum $\xi(f, f_0)$ refers to pairs of sites – *i.e.* it is a 2-SFS – while the spectrum of linked sites $\xi(f|f_0)$ is a single-site SFS.

The relation between both types of spectra can be understood from the relation between the probabilities. The expected spectrum $\mathrm{E}[\xi(f)]$ is given by the probability to find a mutation of frequency $f$ at a specific site, multiplied by the length of the sequence: $\mathrm{E}[\xi(f)] = p(f)L$. As noted above, when $L = 1$ (i.e. a locus with a single site is considered), $\mathrm{E}[\xi(f)]$ corresponds to a proper probability $p(f)$. Assuming the presence of a mutation of frequency $f_0$ at a focal site, we have $\mathrm{E}[\xi(f|f_0)] = p(f|f_0)(L-1)$. For pairs of sites, the expected number of mutations at frequencies $(f, f_0)$ is $\mathrm{E}[\xi(f, f_0)] = p(f, f_0)L(L-1)$ when $f \neq f_0$ or $p(f_0, f_0)L(L-1)/2$ when $f = f_0$. The additional factor $\frac{1}{2}$ accounts for the symmetrical case of equal frequencies $f = f_0$. The equality $p(f, f_0) = p(f|f_0)p(f_0)$ applied to sample and population spectra, results in the following relations:

$$\mathrm{E}[\xi_{k,l}] = \frac{\mathrm{E}[\xi_{k|l}] \cdot \mathrm{E}[\xi_l]}{1 + \delta_{k,l}} = \begin{cases} \mathrm{E}[\xi_{k|l}] \cdot \mathrm{E}[\xi_l] & \text{for } k \neq l \\ \frac{1}{2} \cdot \mathrm{E}[\xi_{k|l}] \cdot \mathrm{E}[\xi_l] & \text{for } k = l \end{cases} \tag{3}$$

$$\mathrm{E}[\xi(f, f_0)] = \frac{\mathrm{E}[\xi(f|f)] \cdot \mathrm{E}[\xi(f)]}{1 + \delta_{f,f_0}} = \begin{cases} \mathrm{E}[\xi(f|f_0)] \cdot \mathrm{E}[\xi(f_0)] & \text{for } f \neq f_0 \\ \frac{1}{2} \cdot \mathrm{E}[\xi(f|f)] \cdot \mathrm{E}[\xi(f)] & \text{for } f = f_0 \end{cases} \tag{4}$$

where $\delta_{x,y}$ is 1 if $x = y$, and 0 otherwise. Note that $x$ and $y$ can be either discrete or continuous variables.

By definition, the 2-SFS includes only pairs of sites that are *both* polymorphic. The probability that a pair of sites contains a single polymorphism of frequency $k/n$ depends only on the 1-SFS and it is approximately equal to $2\mathrm{E}[\xi_k]$ for $\theta \ll 1$. Consequently, on a sequence of size $L$ hosting $S$ polymorphic sites, the number of pairs of sites for which only one of the two is polymorphic of frequency $k/n$ is $\mathrm{E}[(L-S)\xi_k] = L \cdot \mathrm{E}[\xi_k] - \mathrm{E}[S\xi_k] \approx L \cdot \mathrm{E}[\xi_k]$ for small $\theta$.

# 2 Results

## 2.1 Decomposition of the 2-SFS

We follow Sargsyan (2015) and divide the 2-SFS $\xi(f_1, f_2)$ without recombination into two different components: one *nested* component $\xi^N(f_1, f_2)$ for cases where there are individuals carrying the two mutations (one is "nested" in the other), and a *disjoint* component $\xi^D(f_1, f_2)$ that includes disjoint mutations only present in different individuals. The overall spectrum is given by:

$$\xi(f_1, f_2) = \xi^N(f_1, f_2) + \xi^D(f_1, f_2) \tag{5}$$

$$\xi_{k,l} = \xi^N_{k,l} + \xi^D_{k,l} \tag{6}$$

It is noteworthy to mention that that the overall spectrum cannot fully describe the genetic state of the two sites, while the two components $\xi^N(f_1, f_2)$, $\xi^D(f_1, f_2)$ give a complete description up to permutations of all the haplotypes, similarly to the usual SFS for one site. For example, the following haplotypes (derived alleles marked in bold)

$$\begin{array}{ccc} CT & & C\boldsymbol{A} \\ C\boldsymbol{A} & \text{and} & C\boldsymbol{A} \\ \boldsymbol{GA} & & \boldsymbol{G}T \end{array}$$

are identical from the point of view of the overall two-loci spectrum: in both samples there is just a pair of mutations with allele count 1 and 2 respectively, therefore the only (symmetrical) nonzero value of the spectrum is $\xi_{1,2} = \xi_{2,1} = 1$. However the samples can be distinguished by the two components, since in the first one the mutations are nested ($\xi^N_{1,2} = \xi^N_{2,1} = 1$), while in the second one they are disjoint ($\xi^D_{1,2} = \xi^D_{2,1} = 1$). For this reason, these two components constitute the core of the two-loci SFS.

Without recombination, the conditional 1-SFS $\xi(f|f_0)$ can be also decomposed further[2] into different subspectra. They are illustrated in Figure 1:

[2] We subdivide the "strictly nested" mutations of Sargsyan (2015) into *strictly nested* and *enclosing* mutations while we refer to his "identical" mutations as *co-occurring*.

- $\xi^{(sn)}(f|f_0)$ : *strictly nested* mutations, where the mutation is carried only by a subset of individuals with the focal mutation;
- $\xi^{(co)}(f|f_0)$ : *co-occurring* mutations, where both mutations are systematically carried by the same individuals;
- $\xi^{(en)}(f|f_0)$ : *enclosing* mutations, where only a subset of individuals with the mutation also carry the focal one;
- $\xi^{(cm)}(f|f_0)$ : *complementary* mutations, where each individual has only one of the two mutations;
- $\xi^{(sd)}(f|f_0)$ : *strictly disjoint* mutations, where the mutation is carried by a subset of the individuals without the focal one.

Importantly, without recombination, enclosing and complementary mutations cannot be present together in the same sequence.
Given the rules of conditional probabilities $p(f, f_0) = p(f|f_0)p(f_0)$ and the interpretations above, the relations between the two sets of population sub-spectra are:

$$\mathrm{E}[\xi^N(f, f_0)] = \Big(\mathrm{E}[\xi^{(sn)}(f|f_0)] + \mathrm{E}[\xi^{(co)}(f|f_0)] + \mathrm{E}[\xi^{(en)}(f|f_0)]\Big) \cdot \frac{\mathrm{E}[\xi(f_0)]}{1+\delta_{f,f_0}} \tag{7}$$

$$\mathrm{E}[\xi^D(f, f_0)] = \Big(\mathrm{E}[\xi^{(cm)}(f|f_0)] + \mathrm{E}[\xi^{(sd)}(f|f_0)]\Big) \cdot \frac{\mathrm{E}[\xi(f_0)]}{1+\delta_{f,f_0}} \tag{8}$$

Similarly, for sample spectra, we have

$$\mathrm{E}[\xi^N_{k,l}] = \Big(\mathrm{E}[\xi^{(sn)}_{k|l}] + \mathrm{E}[\xi^{(co)}_{k|l}] + \mathrm{E}[\xi^{(en)}_{k|l}]\Big) \cdot \frac{\mathrm{E}[\xi_l]}{1+\delta_{k,l}} \tag{9}$$

$$\mathrm{E}[\xi^D_{k,l}] = \Big(\mathrm{E}[\xi^{(cm)}_{k|l}] + \mathrm{E}[\xi^{(sd)}_{k|l}]\Big) \cdot \frac{\mathrm{E}[\xi_l]}{1+\delta_{k,l}} \tag{10}$$

## 2.2 The joint and conditional SFS

In this section, we report the conditional and joint spectra both for the sample and the population. The derivations and proofs of all equations in this section are given in the Methods and Supplementary Material, as well as comparisons of the analytical spectrum with simulations. The folded version of the 2-SFS is provided in Appendix C for completeness.

### 2.2.1 The sample joint 2-SFS

Using equations 9 and 10, one can derive the two components of the 2-loci spectrum as[3]:

$$\mathrm{E}[\xi^N_{k,l}] = \begin{cases} \theta^2 L^2 \frac{\beta_n(k)-\beta_n(k+1)}{2} & \text{for } k < l \\ \theta^2 L^2 \frac{\beta_n(k)}{2} & \text{for } k = l \\ \theta^2 L^2 \frac{\beta_n(l)-\beta_n(l+1)}{2} & \text{for } k > l \end{cases}$$
$$\mathrm{E}[\xi^D_{k,l}] = \begin{cases} \theta^2 L^2 \left(\frac{1}{kl} - \frac{\beta_n(k)-\beta_n(k+1)+\beta_n(l)-\beta_n(l+1)}{2}\right) \frac{2-\delta_{k,l}}{2} & \text{for } k+l < n \\ \theta^2 L^2 \left(\frac{a_n - a_k}{n-k} + \frac{a_n - a_l}{n-l} - \frac{\beta_n(k)+\beta_n(l)}{2}\right) \frac{2-\delta_{k,l}}{2} & \text{for } k+l = n \\ 0 & \text{for } k+l > n \end{cases} \tag{11}$$

with

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i} \quad , \quad \beta_n(i) = \frac{2n}{(n-i+1)(n-i)}(a_{n+1} - a_i) - \frac{2}{n-i}$$

As shown by equation (6), the full spectrum is simply the sum of the two above equations.

[3] Note that the related formula (14) in the paper by Ferretti *et al.* (2012) has a sign error. It should be identical to the second equation in (11) up to a multiplicative factor.

### 2.2.2 The population joint 2-SFS

Similarly, the 2-SFS for the whole population is given by the sum of the two following equations:

$$\begin{aligned}
\mathrm{E}[\xi^N(f,f_0)] =& \theta^2 L^2 \cdot \left[ \frac{1}{(1-\min(f,f_0))^2} \left( 1 + \frac{1}{\min(f,f_0)} + \frac{2\ln(\min(f,f_0))}{1-\min(f,f_0)} \right) \right. \\
& \left. + \delta(f-f_0) \frac{f_0}{1-f_0} \left( -\frac{\ln(f_0)}{1-f_0} - 1 \right) \right] \\
\mathrm{E}[\xi^D(f,f_0)] =& \theta^2 L^2 \cdot \left[ \frac{1}{ff_0} - \frac{1}{(1-f)^2} \left( 1 + \frac{1}{f} + \frac{2\ln(f)}{1-f} \right) - \frac{1}{(1-f_0)^2} \left( 1 + \frac{1}{f_0} + \frac{2\ln(f_0)}{1-f_0} \right) \right. \\
& \left. + \delta(f-1+f_0) \left( \frac{1-f_0}{f_0^2} \ln(1-f_0) + \frac{f_0}{(1-f_0)^2} \ln(f_0) + \frac{1}{f_0(1-f_0)} \right) \right]
\end{aligned} \tag{12}$$

with $\mathrm{E}[\xi^N(f,f_0)] = 0$ for $f > f_0$ and $\mathrm{E}[\xi^D(f,f_0)] = 0$ for $f + f_0 > 1$. Here, we denote by $\delta(f-f_0)$ the density of the Dirac "delta function" distribution concentrated in $f_0$ (i.e. $\delta(f-f_0) = 0$ for $f \neq f_0$, normalized such as $\int_{-\infty}^{\infty} \delta(f-f_0)df = 1$).

### 2.2.3 The sample conditional 1-SFS

The conditional 1-SFS for sites that are linked to a focal mutation of count $l$ is simply the sum of all its components, given by the following equations:

$$\begin{aligned}
\mathrm{E}[\xi^{(sn)}_{k|l}] =& \theta L \cdot l \frac{\beta_n(k) - \beta_n(k+1)}{2} \quad \text{for } k < l \\
\mathrm{E}[\xi^{(co)}_{k|l}] =& \theta L \cdot l \beta_n(k) \delta_{kl} \\
\mathrm{E}[\xi^{(en)}_{k|l}] =& \theta L \cdot l \frac{\beta_n(l) - \beta_n(l+1)}{2} \quad \text{for } k > l \\
\mathrm{E}[\xi^{(cm)}_{k|l}] =& \theta L \cdot l \left( \frac{a_n - a_k}{n-k} + \frac{a_n - a_l}{n-l} - \frac{\beta_n(k) + \beta_n(l)}{2} \right) \delta_{k,n-l} \\
\mathrm{E}[\xi^{(sd)}_{k|l}] =& \theta L \cdot \left( \frac{1}{k} - l \frac{\beta_n(k) - \beta_n(k+1) + \beta_n(l) - \beta_n(l+1)}{2} \right) \quad \text{for } k + l < n
\end{aligned} \tag{13}$$

and 0 otherwise.

### 2.2.4 The population conditional 1-SFS

For the whole population, this becomes:

$$\begin{aligned}
\mathrm{E}[\xi^{(sn)}(f|f_0)] =& \theta L \cdot \frac{f_0}{(1-f)^2}\left(1+\frac{1}{f}+\frac{2\ln(f)}{1-f}\right), \quad f<f_0 \\
\mathrm{E}[\xi^{(co)}(f|f_0)] =& \theta L \cdot \delta(f-f_0)\frac{2f_0}{1-f_0}\left(-\frac{\ln(f_0)}{1-f_0}-1\right) \\
\mathrm{E}[\xi^{(en)}(f|f_0)] =& \theta L \cdot \frac{f_0}{(1-f_0)^2}\left(1+\frac{1}{f_0}+\frac{2\ln(f_0)}{1-f_0}\right), \quad f>f_0 \qquad (14) \\
\mathrm{E}[\xi^{(cm)}(f|f_0)] =& \theta L \cdot \delta(f-1+f_0)\left[\frac{1-f_0}{f_0}\ln(1-f_0)+\left(\frac{f_0}{1-f_0}\right)^2\ln(f_0)+\frac{1}{1-f_0}\right] \\
\mathrm{E}[\xi^{(sd)}(f|f_0)] =& \theta L \cdot \left[\frac{1}{f}-\frac{f_0}{(1-f)^2}\left(1+\frac{1}{f}+\frac{2\ln(f)}{1-f}\right)\right. \\
&\left.-\frac{f_0}{(1-f_0)^2}\left(1+\frac{1}{f_0}+\frac{2\ln(f_0)}{1-f_0}\right)\right], \quad f<1-f_0
\end{aligned}$$

## 2.3 Shape of the SFS

We report the full joint 2-SFS as well as both the nested and disjoint component (Figure 2). Nested mutations have preferentially a rare mutation in either site – so that the mutation at lower frequency is easily nested into the other – or are co-occurring mutations – corresponding to mutation found in the same branch of the genealogical tree. Disjoint mutations are dominated by cases where both mutations are rare – mostly disjoint – or by complementary mutations. The large contribution of co-occurring (nested component) and complementary mutations (disjoint component) is a direct consequence of the two long branches that coalesce at the root node of a Kingman tree. The conditional 1-SFS of linked sites and the relative contributions of each component to each frequency are shown in Figure 3. Co-occurring and complementary mutations also account for a considerable fraction of the spectrum, especially when the focal mutation ($f_0$) is at high frequency. The rest of the spectrum is biased towards mutations with a lower frequency than the focal one. Strictly nested mutations are important only when the frequency of the focal mutation is intermediate or high. Enclosing mutations are typically negligible and their abundance is uniform as it was also noticed by Hobolth and Wiuf (2009).

Finally, in Figure 4 we show the impact of having a focal mutation of a known frequency on two estimators of $\theta$. The Watterson (1975) estimator, $\hat{\theta}_S$, depends on the total number of polymorphic sites, which increases with the frequency of the focal mutation, while Tajima (1983) estimator, $\hat{\theta}_\pi$, is more sensitive to mutations of intermediate frequency. Therefore the comparison between the two illustrates how the spectrum is skewed towards common or rare mutations. As Tajima's $D$ (Tajima, 1989) is proportional to the difference $\hat{\theta}_\pi - \hat{\theta}_S$, positive values for this test statistic suggest an excess of common mutations while negative values point to an excess of rare mutations. Figure 4 shows that the spectrum has a slight excess of rare mutations at low frequencies of the focal mutation and an excess of common mutations for intermediate frequencies, while it is dominated again by rare mutations if the focal mutation is at high frequencies.

## 2.4 The frequency spectrum of chromosomal inversions

Chromosomal inversions are chromosomal rearrangements in which the orientation of a segment of a chromosome gets reversed. They are well known structural variants, sometimes with important phenotypic effects. Recombination between normal and inverted sequences is strongly suppressed due to mechanical incompatibilities during crossing over, selection against unbalanced chromosomes and presumably other, still unknown, reasons (Kirkpatrick, 2010).

An inversion does not destroy the genetic information of the sequence, but adds a new “allelic” component given by the orientation of the sequence. Apart from the inhibition of recombination, this orientation “allele” is for our purposes akin to a normal point mutation of the same frequency. This is true also for its evolution. Hence, the expected spectrum of neutral inversions can be derived from our results for the linked spectrum, considering the orientation of the sequence as the focal mutation.

If we assume that the original orientation of the sequence is known (e.g. by synteny with a close species) and that the orientation of the sequence is known for all individuals in the sample, then the spectrum of inversions has the same components as the spectrum of sites linked to a focal mutation, as illustrated in Figure 5. This is a consequence of the suppression of recombination between normal and inverted alleles.

We denote the sample spectrum of inversions by $\mathcal{I}_{k|i}$ where $k/n$ is the frequency of mutations and $i/n$ is the frequency of the inversion. If we assume that the rate of inversions is low, i.e. that multiple segregating overlapping inversions are unlikely to occur, then the inversion follows the infinite-sites model. Moreover, recombination within normal or inverted sequences does not affect the joint spectrum of the inversion and a point mutation therein, because it does change not their frequency, nor their linkage. Hence, the expected spectrum of neutral inversions follows directly from our results on the conditional 1-SFS:

$$\mathcal{I}_{k|i}^{(sn)} = \xi_{k|i}^{(sn)}\ ,\ \mathcal{I}_{k|i}^{(co)} = \xi_{k|i}^{(co)}\ ,\ \mathcal{I}_{k|i}^{(en)} = \xi_{k|i}^{(en)}\ ,\ \mathcal{I}_{k|i}^{(cm)} = \xi_{k|i}^{(cm)}\ ,\ \mathcal{I}_{k|i}^{(sd)} = \xi_{k|i}^{(sd)} \quad (15)$$

The same applies to the population spectrum.

If the original orientation of the sequence is unknown, it could be inferred from the frequency spectrum by a Bayesian approach similar to the one employed in Sargsyan (2015) for non-inverted haplotypes.

# 3 Methods

## 3.1 The sample joint 2-SFS

To obtain the sample spectrum for pairs of mutations, we notice that this spectrum can be defined in terms of the expected value of crossproducts of the usual SFS. In detail, we have

$$\mathrm{E}[\xi_{k,l}] = \mathrm{E}[\xi_k \xi_l],\ \text{if } k \neq l \quad (16)$$

and

$$\mathrm{E}[\xi_{k,k}] = \mathrm{E}[\xi_k(\xi_k - 1)]/2. \tag{17}$$

These expected values have been derived by Fu (1995) by coalescent methods. However his results do not distinguish the different contributions from nested and disjoint mutations to the spectrum.

Tracking the origin of each term in the derivation, it is easy to show that equations (24) and (28) of Fu (1995) contribute to nested pairs of mutations, while equations (25), (29) and (30) contribute to disjoint pairs of mutations. All these terms combine linearly and do not interfere, therefore we can decompose the resulting $\mathrm{E}[\xi_k\xi_l]$ into contributions coming from equations (24),(28) and (25),(29) and (30) of Fu (1995). This can be obtain directly by Fu's expression for the covariance matrix $\sigma_{kl}$, since $\mathrm{E}[\xi_k\xi_l] = \delta_{k,l}\mathrm{E}[\xi_k] + \mathrm{E}[\xi_k]\mathrm{E}[\xi_l] + \theta^2 L^2 \sigma_{kl}$ and $\mathrm{E}[\xi_k] = \theta L/k$.

A detailed review of the calculations of Fu (1995), tracking the parts that lead to our mutation classes, is provided in the Supplementary Material.

The same results could also be obtained by re-interpreting the results of Jenkins and Song (2011) from Theorem 5.1 for small $\theta L$ ($\theta$ in their article). Their results for recurrent mutations are mathematically equivalent to the results for mutations in an infinite-sites model, for a special choice of allele transition matrices (in the triallelic case, a strictly lower triangular matrix with all non-zero entries equal to 1). Their classification is based on the location of the mutations on the tree: their "nested mutations" correspond to strictly nested and enclosing mutations here, "mutations on the same branch" correspond to co-occurring mutations, "mutations on basal branches" correspond to complementary mutations, and "non-nested mutations' correspond to strictly disjoint mutations.

## 3.2 The sample conditional 1-SFS

The spectrum for sites linked to a focal mutation of count $l$ (equation 13) can be obtained from the previous spectrum (11). The first step is simply to condition on the frequency $l/n$ of the focal mutation, i.e. dividing the

2-SFS $\mathrm{E}[\xi_{k,l}]$ by $\mathrm{E}[\xi_l]\frac{1+\delta_{k,l}}{2}$ following equations (9) and (10). In fact, $\mathrm{E}[\xi_{k|l}] = (L-1)\mathrm{P}[c(x) = k|c(y) = l] = L(L-1)\mathrm{P}[c(x) = k, c(y) = l]/L\mathrm{P}[c(y) = l] = \frac{2}{1+\delta_{k,l}}\mathrm{E}[\xi_{k,l}]/\mathrm{E}[\xi_l]$ where $c(x)$ is the derived allele count at site $x$.

The second step is to break further the two contributions of the resulting conditional spectrum into the different components. Strictly nested, co-occurring and enclosing mutations are derived from the nested contribution and are distinguished by site frequencies only: strictly nested ones correspond to $k < l$, co-occurring ones to $k = l$ and enclosing ones to $k > l$. Similarly, from the disjoint contribution, mutations belonging to the strictly disjoint component can be obtained by selecting the frequency range $k + l < n$ while complementary ones correspond to $k + l = n$.

### 3.3 Population spectra

In the limit of large samples, the frequency spectra converge to the continuous SFS for infinite populations. However, the limit $n \to \infty$ should be taken with care. The easiest derivation proceeds as follows: since the conditional 1-SFS (eq 14) is a single-locus spectrum, its population components can be obtained from the corresponding ones for finite samples (eq. 13) by direct application of the equation (2). Then the population 2-SFS (eq 12) can be reconstructed from equations (7) and (8), by multiplying by the neutral spectrum $\mathrm{E}[\xi(f_0)] = \theta L/f_0$ and by $\frac{1}{1+\delta_{f,f_0}}$ and combining the result into nested and disjoint contributions. The only tricky passage of the derivation is the following functional limit of the Kronecker delta as a Dirac delta function: $n\delta_{\lfloor nf \rfloor, \lfloor nf_0 \rfloor} \to \delta(f - f_0)$ for $n \to \infty$. More details are given in the Supplementary Material.

## 4 Discussion

In this article, we have provided the first exact closed formulae for the joint 2-SFS as well as for the conditional 1-SFS, both for sample and population. Using the basic results from Fu (1995), we were able to derive the formulae for sample spectra which we used then to derive the population spectra by letting $n \to \infty$. Importantly, our results only hold when there is no recombination, and are averaged across the tree space.

The analytical expressions provided in this paper can be intuitively understood in terms of the evolution of linked mutations. Consider a new mutation increasing in frequency by neutral drift and reaching low/intermediate frequency. We expect to find a large number of strictly disjoint and a low number of strictly nested linked mutations, since at the time of appearance of the focal mutation all other mutations were "strictly disjoint". Enclosing mutations are more abundant than strictly nested, but less or abundant as strictly disjoint mutations, depending on the initial frequency of the focal mutation. The spectrum of strictly nested mutations is more skewed towards rare alleles than predicted by the neutral spectrum $1/f$, since strictly nested mutations evolve inside an expanding subpopulation. On the other hand, the spectrum of strictly disjoint mutations resembles the neutral one but with a slight bias against rare mutations, since they evolved in a slightly contracting subpopulation.

Note that for sequences linked to a mutation close to fixation, co-occurring and complementary mutations dominate. The contrast between the haplotypes produces a strong "haplotype structure".

Interestingly, conditioning on the presence of a mutation of frequency $f$ impacts the length and balance of the coalescent, as apparent from Figure 4. This can be understood as follows. Rare mutations are common in any realisation of the coalescent tree but especially common in the lower branches, therefore they just increase slightly the tree length and the length of the lower branches compared to the unconditioned case. Instead, mutations of intermediate frequency appear mostly in the upper branches of the tree, therefore the presence of such mutations implies higher, more balanced trees. The effect is even stronger for high frequency mutations, which reside only in the uppermost branches, implying high unbalanced trees.

There are several potential applications of these results. Direct applications include the improvement of population genetic inference techniques based on the SFS, such as composite likelihood (*e.g.,* Kim and Stephan, 2002; Li and Stephan, 2005; Kim and Nielsen, 2004; Nielsen *et al.*, 2005) and Poisson Random Field methods (Sawyer and Hartl, 1992). These methods use analytical expressions for the SFS for a single site together with approximations of independence between different sites. For sequences with low recombination, they could be made more rigorous by assuming independence between different *pairs* of sites, while taking pairwise dependence between sites into account through the two-locus SFS developed here.

The spectrum could also be useful for new neutrality tests based on linkage between mutations. Our results lead to a better understanding of the linkage disequilibrium (LD) structure among neutral loci, therefore they can be immediately applied to LD-related statistics, for example to compute average LD across non-recombining neutral loci. Furthermore, they can be used to build neutrality tests optimised to detect positive or balancing selection through its effect on the frequency spectrum of linked sites.

An example of direct application of our results is the spectrum of chromosomal inversions and other structural variants. These genomic variants often have phenotypic effects and their evolutionary dynamics is of significant interest. We provided the spectrum of the null neutral model, that is a fundamental step to build methods for the detection of non-neutral evolution. Further work on the derivation of appropriate neutrality tests, their optimisation and application will be presented in future publications.

The spectra presented here could also provide a neutral model for other scenarios, including introgressions from different species or populations. Our results contribute to the ongoing search for genetic signatures of selection on introgressed alleles.

The SFS presented here is the simplest two-locus spectrum for neutral, non-recombining mutations in a population of constant size. These results could be extended to variable population size using the approach of Živković and Wiehe (2008); Jenkins and Song (2011) and to mutations in rapidly adapting populations using the Λ-coalescent approximation and the results of Birkner *et al.* (2013). However, the most interesting extensions would be to consider (a) non-neutral mutations and (b) recombination.

Adding selection to the two-locus SFS would significantly enhance its potential for most of the applications discussed above. The SFS for pairs of selected mutations has been obtained by Xie (2011) as a polynomial expansion. However, the computation is still cumbersome, while flexible numerical alternative could be soon available. Given the simplicity of the expression for the single-locus SFS $\xi(f) = \theta(1-e^{-2N_e s(1-f)})/f(1-f)(1-e^{-2N_e s})$ (Wright, 1938; Sawyer and Hartl, 1992), we expect that closed expressions could be found for pairs of mutations with different selective coefficients. This would be a promising development for future investigations.

The classical correspondence between the Kingman model in the large $n$ limit and the diffusion approximation suggests that the 2-SFS spectrum presented here is a solution of the diffusion equations for three alleles (Ewens, 2012). In fact, it is easy to check that the nested component of the 2-SFS for $f \neq f_0$ is a stationary solution of the diffusion equation of three alleles of frequency $f$, $f_0 - f$ and $1 - f_0$:

$$\frac{\partial \xi}{\partial t} = \frac{1}{2N_e}\left(\frac{\partial^2}{\partial f^2}\left[f(1-f)\xi\right] + 2\frac{\partial^2}{\partial f \partial f_0}\left[f(1-f_0)\xi\right] + \frac{\partial^2}{\partial f_0^2}\left[f_0(1-f_0)\xi\right]\right) \tag{18}$$

while the disjoint component for $f \neq 1 - f_0$ is a stationary solution of the diffusion equation of three alleles of frequency $f$, $f_0$ and $1 - f_0 - f$:

$$\frac{\partial \xi}{\partial t} = \frac{1}{2N_e}\left(\frac{\partial^2}{\partial f^2}\left[f(1-f)\xi\right] - 2\frac{\partial^2}{\partial f \partial f_0}\left[f f_0 \xi\right] + \frac{\partial^2}{\partial f_0^2}\left[f_0(1-f_0)\xi\right]\right) \tag{19}$$

The correspondence implies that the solution (12) is actually the stationary solution of the full set of diffusion equations for the system, including boundary equations for $f = f_0$ and $1 - f_0$ and boundary conditions. A direct proof of this result using methods from the theory of partial differential equations could lead to interesting developments towards new solutions for selective equations as well.

On the other hand, finding the exact two-locus SFS with recombination appears to be a difficult problem. Recombination is intrinsically related to the two-locus SFS via the same definition of linkage disequilibrium. Obtaining the full two-locus spectrum with selection and recombination could open new avenues for model inference and analysis of genomic data. For this reason, many approximations and partial results have been developed since Hudson (2001), like expansions in the limit of strong recombination (Jenkins and Song, 2012). The SFS of linked loci presented in this paper could be useful as a starting point for different approaches to the effect of recombination events, for example for perturbation expansions at low recombination rates. There is actually an immediate application of our results to recombination events. Since in the Ancestral Recombination Graph (Griffiths and Marjoram, 1997) the recombination events follow a Poisson process similar to mutation events, although with a different rate, the spectrum $\xi_{k|l}$ could also be reinterpreted (up to a constant) as the probability that a single *recombination* event affects $k$ extant lineages in a sequence linked to a specific mutation of frequency $l$, i.e. it is equivalent to the spectrum of mutation-recombination events. This approach could be applied to higher moments of the frequency spectrum and lead to new results in recombination theory.

# Acknowledgments

We thank Wolfgang Stephan for insightful discussions. GA and LF were supported by grant ANR-12-JSV7-0007 TempoMut from Agence Nationale de la Recherche. GA was also supported by grant ANR-12-BSV7-0012 Demochips, AK and TW by grants of the German Science Foundation (DFG-SFB680 and DFG-SPP 1590).

# References

Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics **183**: 249–58.

Alcala, N., J. D. Jensen, A. Telenti, and S. Vuilleumier, 2016 The genomic signature of population reconnection following isolation: From theory to hiv. G3: Genes— Genomes— Genetics **6**: 107–120.

Birkner, M., J. Blath, and B. Eldon, 2013 Statistical properties of the site-frequency spectrum associated with lambda-coalescents. Genetics : genetics–113.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, *et al.*, 2002 The cost of inbreeding in arabidopsis. Nature **416**: 531–4.

Bustamante, C. D., J. Wakeley, S. Sawyer, and D. L. Hartl, 2001 Directional selection and the site-frequency spectrum. Genetics **159**: 1779–88.

Corbett-Detig, R. B., and D. L. Hartl, 2012 Population genomics of inversion polymorphisms in drosophila melanogaster. PLoS Genet **8**: e1003056.

Durrett, R., 2008 *Probability models for DNA sequence evolution*. Springer.

Ethier, S., and R. Griffiths, 1990 On the two-locus sampling distribution. Journal of Mathematical Biology **29**: 131–159.

Ewens, W. J., 2012 *Mathematical Population Genetics 1: Theoretical Introduction*, volume 27. Springer.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive darwinian selection. Genetics **155**: 1405–13.

Ferretti, L., M. Perez-Enciso, and S. Ramos-Onsins, 2010 Optimal neutrality tests based on the frequency spectrum. Genetics **186**: 353–65.

Ferretti, L., E. Raineri, and S. Ramos-Onsins, 2012 Neutrality tests for sequences with missing data. Genetics **191**: 1397–1401.

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

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

Golding, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108**: 257–74.

Griffiths, R., 1979 A transition density expansion for a multi-allele diffusion model. Advances in Applied Probability : 310–325.

Griffiths, R., and S. Tavare, 2003 The genealogy of a neutral mutation. Oxford Statistical Science Series : 393–413.

Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph. Institute for Mathematics and its Applications **87**: 257.

Griffiths, R. C., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci **344**: 403–10.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS Genet **5**: e1000695.

Hobolth, A., and C. Wiuf, 2009 The genealogy, site frequency spectrum and ages of two nested mutant alleles. Theoretical population biology **75**: 260–265.

Hoffmann, A. A., C. M. Sgrò, and A. R. Weeks, 2004 Chromosomal inversion polymorphisms and adaptation. Trends in Ecology & Evolution **19**: 482–488.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159**: 1805–1817.

Hudson, R. R., 2002 Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18**: 337–8.

Hudson, R. R., *et al.*, 1990 Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology **7**: 44.

Jenkins, P. A., J. W. Mueller, and Y. S. Song, 2014 General triallelic frequency spectrum under demographic models with variable population size. Genetics **196**: 295–311.

Jenkins, P. A., and Y. S. Song, 2011 The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. Theoretical population biology **80**: 158–173.

Jenkins, P. A., and Y. S. Song, 2012 Padé approximants and exact two-locus sampling distributions. The Annals of Applied Probability **22**: 576–607.

Kennington, W. J., L. Partridge, and A. A. Hoffmann, 2006 Patterns of diversity and linkage disequilibrium within the cosmopolitan inversion in (3r) payne in drosophila melanogaster are indicative of coadaptation. Genetics **172**: 1655–1663.

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167**: 1513–24.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160**: 765–77.

Kimura, M., 1956 Random genetic drift in a tri-allelic locus; exact solution with a continuous model. Biometrics **12**: 57–66.

Kimura, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, Great Britain.

Kingman, J. F. C., 1982 The coalescent. Stochastic processes and their applications **13**: 235–248.

Kirkpatrick, M., 2010 How and Why Chromosome Inversions Evolve. PLoS Biology **8**: e1000501.

Krimbas, C. B., and J. R. Powell, 1992 *Drosophila inversion polymorphism*. CRC Press.

Ledda, A., G. Achaz, T. Wiehe, and L. Ferretti, 2015 Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. arXiv preprint arXiv:1510.06748 .

Li, H., and W. Stephan, 2005 Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. Genetics **171**: 377–84.

Littler, R., and E. Fackerell, 1975 Transition densities for neutral multi-allele diffusion models. Biometrics : 117–123.

Liu, X., and Y.-X. Fu, 2015 Exploring population size changes using snp frequency spectra. Nature genetics **47**: 555–559.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, *et al.*, 2005 Genomic scans for selective sweeps using snp data. Genome Res **15**: 1566–75.

Sargsyan, O., 2015 An analytical framework in the general coalescent tree setting for analyzing polymorphisms created by two mutations. J Math Biol **70**: 913–56.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132**: 1161–1176.

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

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics **123**: 585–95.

Thornton, K., 2005 Recombination and the properties of tajima's d in the context of approximate-likelihood calculation. Genetics **171**: 2143–8.

Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical population biology **7**: 256–276.

White, B. J., C. Cheng, D. Sangaré, N. F. Lobo, F. H. Collins, *et al.*, 2009 The population genomics of trans-specific inversion polymorphisms in anopheles gambiae. Genetics **183**: 275–288.

Wright, S., 1938 The distribution of gene frequencies under irreversible mutation. Proceedings of the National Academy of Sciences of the United States of America **24**: 253.

Xie, X., 2011 The site-frequency spectrum of linked sites. Bulletin of mathematical biology **73**: 459–494.

Živković, D., and T. Wiehe, 2008 Second-order moments of segregating sites under variable population size. Genetics **180**: 341–357.

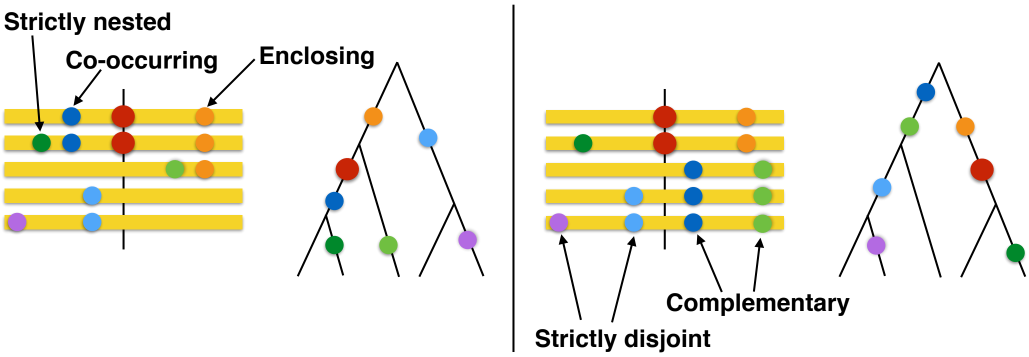

Figure 1: An illustration of two non-recombining loci and their corresponding genealogical trees. The yellow segments represent the ancestral sequence and the colored bullets represent derived alleles. This figure illustrates the classification of all possible types of mutations with respect to the focal mutation (in red) and their occurrence on the sequence tree. Nested mutations are indicated in the left panel, disjoint mutations in the right one.

If the focal mutation is not on a root branch (left), it is clear from the figures that mutations can be on the same branch as the focal mutation (*co-occurring*), on the subtree below (*strictly nested*), between the focal mutation and the root (*enclosing*), or on other branches (*strictly disjoint*). If the mutation is on a root branch (right), there cannot be enclosing mutations, but there are mutations on the other root branch (*complementary*).

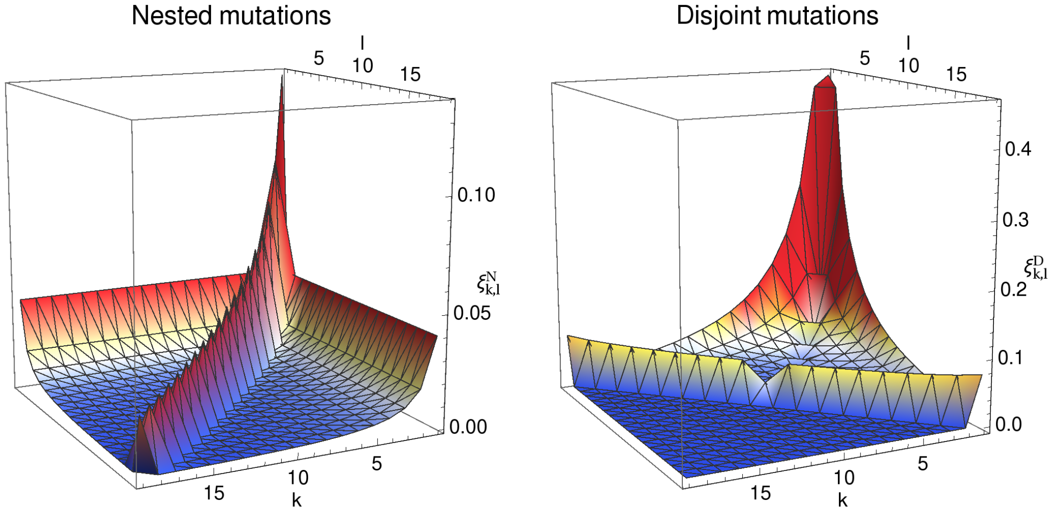

Figure 2: Plots of nested and disjoint contributions to the two-locus frequency spectrum for $\theta L = 1$, $n = 20$. Note the different scales of the two plots.

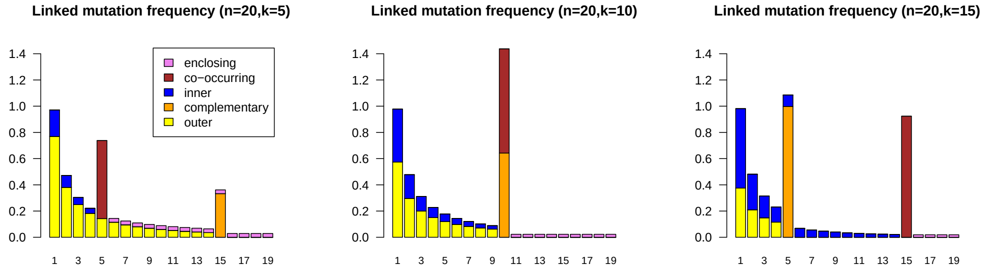

Figure 3: Barplot of the spectrum of linked sites for $\theta L = 1$, $n = 20$, each column colored according to the different contributions. The focal mutation has frequency 5/20=0.25 (left), 10/20=0.5 (middle) and 15/20=0.75 (right) respectively.

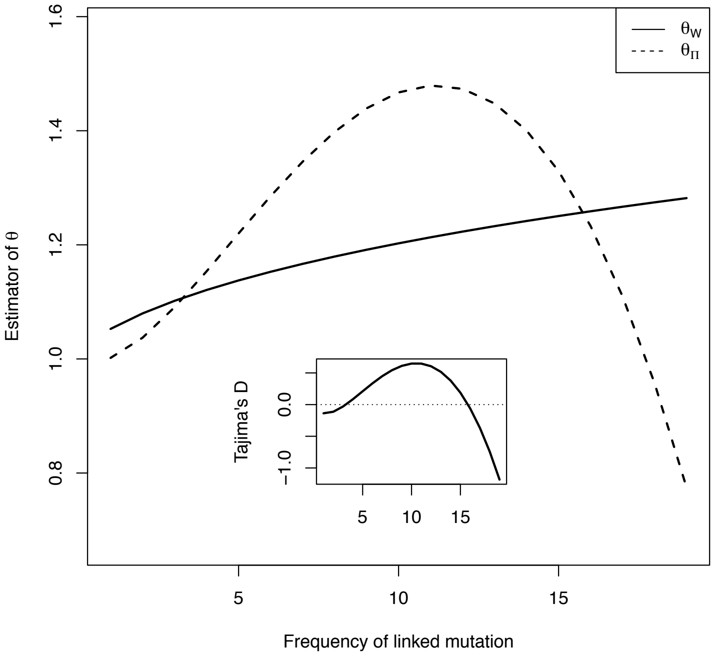

Figure 4: Mean values of the Watterson estimator ($\hat{\theta}_S$) and Tajima estimator ($\hat{\theta}_\pi$) of $\theta$ conditioned on the presence of a linked mutation, for $\theta = 1$, $n = 20$. In the inset, approximate mean value of Tajima's $D$ (computed substituting $S$ with its mean value in the denominator).

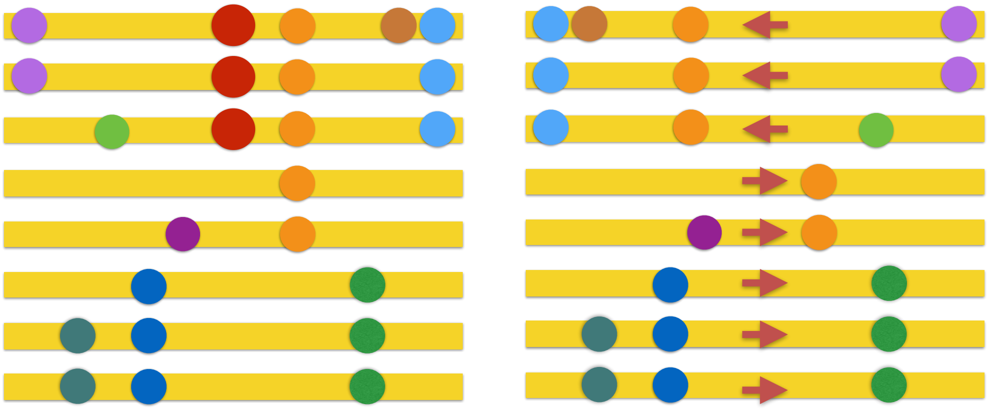

Figure 5: Illustration of the similarity between inversions (right) and SNPs (left). The yellow segments represent the ancestral sequence, the red arrows represent its orientation, while the colored bullets represent derived SNPs.

# A 2-SFS for ordered pairs of sites

The expected spectrum of linked sites described in the previous sections applies to unordered pairs of sites. As an example, consider a sequence containing just two nested SNPs with mutations of frequency 0.3 and 0.1 respectively. The nonzero components of the spectrum are $\xi(0.1, 0.3) = \xi(0.3, 0.1) = 1$, irrespective of which of the two SNPs has frequency 0.1.

However, it can be useful to rewrite our results in terms of the spectrum $\xi^{ordered}$ for ordered pairs of sites. Sites can be ordered by their position along the sequence, or by any other criterion. In the previous example, the components of the ordered spectrum are $\xi^{ordered}(0.3, 0.1) = 1$ but $\xi^{ordered}(0.1, 0.3) = 0$.

The relation between the 2-SFS and the ordered 2-SFS is the following. For different frequencies $k \neq l$, the 2-SFS of unordered pairs is symmetric, so $\xi_{k,l} = \xi_{l,k}$ are actually the same object. However, for the ordered 2-SFS, they are different. Their sum correspond to the total number of unordered pairs:

$$\xi_{k,l} = \xi_{k,l}^{ordered} + \xi_{l,k}^{ordered} \tag{20}$$

and since the expected values do not depend on the order,

$$\mathrm{E}[\xi_{k,l}^{ordered}] = \mathrm{E}[\xi_{l,k}^{ordered}] = \mathrm{E}[\xi_{k,l}]/2 \tag{21}$$

On the other hand, the order does not matter for pairs of identical mutations, i.e. $\xi_{k,k} = \xi_{k,k}^{ordered}$ and therefore

$$\mathrm{E}[\xi_{k,k}^{ordered}] = \mathrm{E}[\xi_{k,k}] \tag{22}$$

These relations can be extended to the population spectrum in a straightforward way. Note that this factor of 2 between both cases relates to the same factor in equations (4) and (3). In fact, $\mathrm{E}[\xi_{k,l}^{ordered}] = \mathrm{E}[\xi_{k|l}] \cdot \mathrm{E}[\xi_l]$.

# B Triallelic spectrum

As discussed before, the evolutionary dynamics of two non-recombining SNPs is the same as the one of a triallelic locus, where the three alleles are represented by the possible haplotypes of the sequence containing the SNPs. Therefore we can extract the frequency spectrum of neutral mutations in a triallelic non-recombining locus from our results.

Triallelic loci can represent many possible types of variants in genomes. Triallelic SNPs can be present in any set of nucleotide sequences - however these sites are rare compared to biallelic SNPs. Or they could be Copy Number Variants, or microsatellites with variable number of repeats.

The unfolded tri-allelic spectrum for two derived alleles of frequency $f_1, f_2$ generated with rescaled mutation rates per locus $\theta_1^{loc}, \theta_2^{loc}$ is

$$\mathrm{E}[\xi^{3al}(f_1, f_2)] = \theta_1\theta_2 \left(\mathrm{E}[\xi^D(f_1, f_2)] + \mathrm{E}[\xi^N(f_2 - f_1, f_1)] + \mathrm{E}[\xi^N(f_1 - f_2, f_2)]\right), \tag{23}$$

where the expectations are given by equation (12) with $\theta = 1$.

Similarly, the sample triallelic spectrum for derived alleles of count $k, l$ is

$$\mathrm{E}[\xi^{3al}_{k,l}] = \theta_1\theta_2 \left(\mathrm{E}[\xi^D_{k,l}] + \mathrm{E}[\xi^N_{k-l,l}] + \mathrm{E}[\xi^N_{l-k,k}]\right), \tag{24}$$

where the expectations are given by equation (11) with $\theta = 1$. This spectrum was also derived by Jenkins and Song (2011) for a general matrix of mutation rates.

# C The folded spectra

When no reliable outgroup sequence is available, one cannot assess if the allele is derived or ancestral. In that case, alleles can only be classified as minor (less frequent) and major (most frequent). The distribution of minor allele frequencies, known as the folded SFS, will be noted $\eta(f^*)$, where $f^*$ denotes the minor allele frequency that ranges from 0 to 0.5. Importantly, the folded SFS can be retrieved from the full SFS by simply summing alleles at complementary frequencies:

$$\eta(f^*) = [\xi(f^*) + \xi(1 - f^*)]/(1 + \delta_{f^*,(1-f^*)}) \tag{25}$$

As a consequence, the single site SFS under the standard neutral model then become $\mathrm{E}[\eta(f^*)] = \theta/[f^*(1-f^*)(1+\delta_{f^*,(1-f^*)})]$ and $\mathrm{E}[\eta_{k^*}] = \theta n/[k^*(n-k^*)(1+\delta_{k^*,n-k^*})]$, where $k^*$ denotes the count of the minor allele.

Following the same idea, we define a conditional folded 1-SFS and a joint folded 2-SFS using the minor allele frequencies. Minor alleles can also be classified as "nested" or "disjoint" depending on the presence or absence of individuals enclosing both minor alleles. As for the unfolded case, this classification gives a complete description of the linkage between pairs of mutations. However, in contrast to the unfolded case, the classification has no strict evolutionary meaning. For example, "disjoint" minor alleles do not necessarily correspond to pairs of alleles born in different backgrounds. Moreover, alleles of frequency $f^* = 0.5$ (or allele count $k^* = n/2$) suffer from an ambiguity in the choice of the minor allele and therefore should be treated separately. Note also that with the exception of alleles with frequency 0.5, folded spectra do not contain complementary alleles, since the frequency of one of the two complementary alleles will exceed 0.5.

Pairs of mutations with $f, f_0$ both larger or smaller than 0.5 will be classified identically (as nested or disjoint) in the folded case. However, pairs of mutations with $f < 0.5$ and $f_0 > 0.5$ (or vice-versa) will swap their classification. As a consequence, the two components of the 2-SFS are:

$$\begin{aligned}\mathrm{E}[\eta^N(f^*, f_0^*)] =& \mathrm{E}[\xi^N(f^*, f_0^*)] + \mathrm{E}[\xi^N(1-f^*, 1-f_0^*)] + \mathrm{E}[\xi^D(f^*, 1-f_0^*)] \\ & + \mathrm{E}[\xi^D(1-f^*, f_0^*)] \\ \mathrm{E}[\eta^D(f^*, f_0^*)] =& \mathrm{E}[\xi^D(f^*, f_0^*)] + \mathrm{E}[\xi^N(f^*, 1-f_0^*)] + \mathrm{E}[\xi^N(1-f^*, f_0^*)] \end{aligned} \quad (26)$$

To obtain the conditional 1-SFS, we proceed similarly to the unfolded case. First we separate the 2-SFS above into components based on frequency. The strictly nested component corresponds to frequencies $f^* < f_0^*$ of the nested part, while the cooccurring and enclosing components corresponds to $f^* = f_0^*$ and $f^* > f_0^*$ respectively. The strictly disjoint component corresponds to the disjoint part, since there cannot be any complementary component. Then we divide each component by the expected 1-SFS $\mathrm{E}[\eta(f_0^*)]$ to obtain

$$\begin{aligned}
\mathrm{E}[\eta^{(sn)}(f^*|f_0^*)] =&\frac{f_0^*(1-f_0^*)}{\theta}\mathrm{E}[\eta^N(f^*,f_0^*)] \quad \text{for } f^* < f_0^* \\
\mathrm{E}[\eta^{(co)}(f^*|f_0^*)] =&2\cdot\frac{f_0^*(1-f_0^*)}{\theta}\mathrm{E}[\eta^N(f^*,f_0^*)] \quad \text{for } f^* = f_0^* \\
\mathrm{E}[\eta^{(en)}(f^*|f_0^*)] =&\frac{f_0^*(1-f_0^*)}{\theta}\mathrm{E}[\eta^N(f^*,f_0^*)] \quad \text{for } f^* > f_0^* \\
\mathrm{E}[\eta^{(cm)}(f^*|f_0^*)] =&0 \\
\mathrm{E}[\eta^{(sd)}(f^*|f_0^*)] =&(1+\delta_{f^*,f_0^*})\cdot\frac{f_0^*(1-f_0^*)}{\theta}\mathrm{E}[\eta^D(f^*,f_0^*)]
\end{aligned} \tag{27}$$

While the classification of the pairs with frequencies $f^* = 0.5$ and/or $f_0^* = 0.5$ is ambiguous, these pairs are usually irrelevant for the population spectrum. The sample spectra are similar. For $n$ even, there are ambiguous pairs with $k$ or $l = n/2$ that can be easily retrieved from the equations (11),(13) and treated separately. Considering only $k, l < n/2$, the sample 2-SFS is:

$$\begin{aligned}
\mathrm{E}[\eta^N_{k^*,l^*}] =&\mathrm{E}[\xi^N_{k^*,l^*}] + \mathrm{E}[\xi^N_{n-k^*,n-l^*}] + \mathrm{E}[\xi^D_{k^*,n-l^*}] + \mathrm{E}[\xi^D_{n-k^*,l^*}] \\
\mathrm{E}[\eta^D_{k^*,l^*}] =&\mathrm{E}[\xi^D_{k^*,l^*}] + \mathrm{E}[\xi^N_{k^*,n-l^*}] + \mathrm{E}[\xi^N_{n-k^*,l^*}]
\end{aligned} \tag{28}$$

and the conditional 1-SFS is:

$$\begin{aligned}
\mathrm{E}[\eta^{(sn)}_{k^*|l^*}] =&\frac{l^*(n-l^*)}{\theta n}\mathrm{E}[\eta^N_{k^*,l^*}] \quad \text{for } k^* < l^* \\
\mathrm{E}[\eta^{(co)}_{k^*|l^*}] =&2\cdot\frac{l^*(n-l^*)}{\theta n}\mathrm{E}[\eta^N_{k^*,l^*}] \quad \text{for } k^* = l^* \\
\mathrm{E}[\eta^{(en)}_{k^*|l^*}] =&\frac{l^*(n-l^*)}{\theta n}\mathrm{E}[\eta^N_{k^*,l^*}] \quad \text{for } k^* > l^* \\
\mathrm{E}[\eta^{(cm)}_{k^*|l^*}] =&0 \\
\mathrm{E}[\eta^{(sd)}_{k^*|l^*}] =&(1+\delta_{k^*,l^*})\cdot\frac{l^*(n-l^*)}{\theta n}\mathrm{E}[\eta^D_{k^*,l^*}]
\end{aligned} \tag{29}$$

# Supplementary Material

## S.1 Classification of two linked mutations

As discussed also by Sargsyan (2015), it is easy to see that our mutation classes cover all possible relations of two mutations in a non-recombining coalescent. The two bi-allelic sites were created by two independent mutations: an *old* mutation followed by a *young* one. They both occurred in a single individual and then rose in frequency throughout the action of genetic drift. The young mutation could have occurred in an individual that also carried the old mutation, leading to what we have name the "nested" case.

Conversely, if the young mutation has occurred in an individual who did not have the old mutation, it leads to the "disjoint" case. As recombination is forbidden here, the complete linkage prevents any further mixing between these two cases and the derived allele that corresponds to the young mutation will remain fully linked to the background allele it occurred in.

In the nested case, the young mutation can be fixed in sequences carrying the old one (that is co-occurring case) or not. In the latter case, the young mutation can be the focal one (enclosing case) or the other one (strictly nested case).

In the disjoint case, the young mutation can get fixed among the individuals lacking the old mutation (complementary case) or not (strictly disjoint case). Therefore, without recombination, these 5 types are the only possible cases. Because these are the only possible classes of mutations without recombination, there are constrains on the frequency spectrum for linked sites. For example, the presence of an enclosing mutation of count $k$ is incompatible with complementary mutations or strictly disjoint mutations of count greater than $n - k$; this can be shown by considering the enclosing mutation as focal one, and noticing that the other mutations would not fall in any of the previous classes.

## S.2 Simulations

In this section we present a numerical result as an example to check the consistency of our results. In Figure S1 the analytical sample spectrum is compared with those obtained by coalescent simulations. We parsed the output of *ms* (Hudson, 2002) to count the number of mutations conditional on a focal mutation of given frequency. The good agreement between the spectra supports our equations.

The source code (C++) for computing analytical as well as simulated spectra can be found in the package *coatli* developed by one of the authors and available on `http://sourceforge.net/projects/coatli/`.

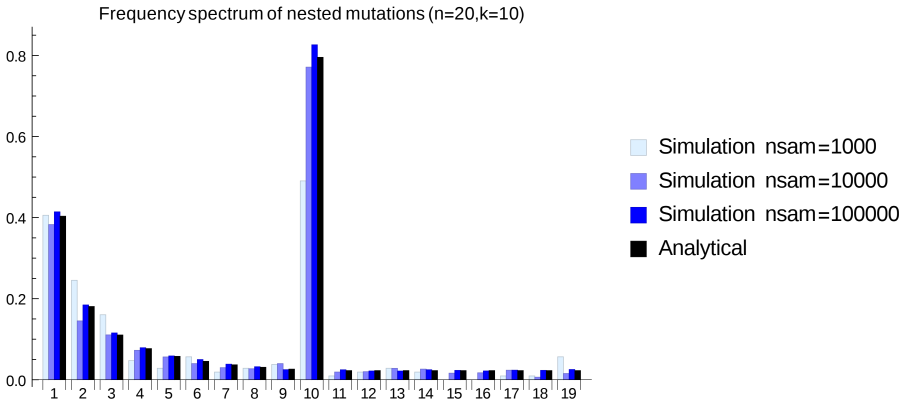

Figure S1: Frequency spectrum of nested mutations in linked sites for n=20, $L\theta = 1$ and a focal mutation of frequency $k = 10$, compared with coalescent simulations (averages for different numbers of samples).

## S.3 Derivation of the sample 2-SFS: Fu (1995) reloaded

The 1995 paper by Fu (1995) derived the second moments of the Kingman coalescent (Kingman, 1982), more precisely the covariance of mutations of size $i$ and $j$: $\mathrm{Cov}[\xi_i, \xi_j]$. Unfortunately the very tight presentation and some typos may make it hard to follow the transformations. A valuable introduction into the proof, using a different notation, has been given in Durrett

(2008), omitting the more technical parts. Since the latter are important for us, we reproduce the essential parts of the proof in greater detail and original notation, and show how they lead to our expressions for different mutations classes.

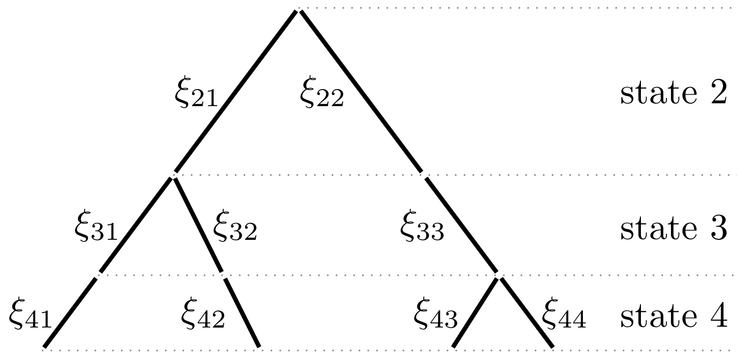

Figure S2: A coalescent tree describing the genealogy of a non-recombining locus for a sample of size $n = 4$. The topology of the tree is defined by the relationship between the lines, e.g. line $\xi_{43}$ is a descendant of lines $\xi_{33}$ and $\xi_{22}$, but not of any other line. A mutation happening "on" line $\xi_{33}$ is of size 2, since it has two descendant lines (and hence leaves) at state $n = 4$, i.e. two individuals of the sample carry it. All lines of the same state have the same length, reflecting the same mutation probability. Hence the amount of mutations of size 1 ("singletons") occurring on $\xi_{31}$ and $\xi_{32}$ is correlated with the amount of mutations of size 2 arising on $\xi_{33}$. Averaging over different topologies leads to more complicated correlations.

As a starting point for the combinatorics let us note that the descendance of lines in the coalescent can be described by a Polya urn process, and the two expressions given beneath are special cases of a general formula (c.f. e.g. Griffiths and Tavare (2003)). We introduce the following notation: let $p_{k\to n}(t \to i)$ denote the probability that $t$ lines at state $k$ have $i$ descendents at state $n$. This probability is

$$p_{k\to n}(t\to i)=\frac{\binom{i-1}{t-1}\binom{n-i-1}{k-t-1}}{\binom{n-1}{k-1}}$$

and the probability that $t$ and $u$ lines at state $k$ have respectively $i$ and $j$ descendents at state $n$ is

$$p_{k\to n}(t\to i,u\to j)=\frac{\binom{i-1}{t-1}\binom{j-1}{u-1}\binom{n-i-j-1}{k-t-u-1}}{\binom{n-1}{k-1}}\ .$$

In order to avoid case distinctions it is helpful to abuse for a while the notation by defining $\binom{-1}{-1}=1$ and $\binom{n}{k}=0$ for any other combination of $n<0$ or $k<0$ (as has been employed by DURRETT (2008), too). This makes it possible to subsume in the above and following formulas “boundary cases” such as $k$ lines of state $k$ yielding the $n$ lines of state $n$ (with probability 1). Later on these special cases will be considered separately and the final expressions don’t contain any negative values.

The probability that a line at state $k$ is of size $i$ is referred to as $p(k,i)$. The probability that two lines at state $k$ are of size $i$ and $j$ is referred to as $p(k,i;k,j)$. The probability that a line at state $k$ and another at state $k'>k$ are of size $i$ respective $j$ is split up with respect to the latter line being a descendant of the former line or not: $p(k,i;k',j)=p_a(k,i;k',j)+p_b(k,i;k',j)$.

The two formulas above suffice to derive these probabilities:

$$p(k,i) = p_{k\to n}(1\to i) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}}$$

$$p(k,i;k,j) = p_{k\to n}(1\to i, 1\to j) = \frac{\binom{n-i-j-1}{k-3}}{\binom{n-1}{k-1}}$$

$$\begin{aligned}
p_a(k,i;k',j) &= \sum_{t=1}^{k'-1} p_{k\to k'}(1\to t)\frac{t}{k'}p_{k'\to n}(1\to j, t-1\to i-j) \\
&= \sum_{t=1}^{k'-1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{t}{k'} \frac{\binom{i-j-1}{t-2}\binom{n-i-1}{k'-t-1}}{\binom{n-1}{k'-1}} \\
p_b(k,i;k',j) &= \sum_{t=1}^{k'-1} p_{k\to k'}(1\to t)\frac{k'-t}{k'}p_{k'\to n}(1\to j, t\to i) \\
&= \sum_{t=1}^{k'-1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{k'-t}{k'} \frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}}
\end{aligned}$$

In the latter two formulas, the summation index $t$ stands for the number of descendants, that the line from state $k$ may have at state $k'$.

Now we consider the "mutational" correlation between the lines. Other than in our main article, $\theta$ denotes here the *locus* mutation rate (not the site mutation rate), i.e. includes the locus length $L$.

Let $X$ be a random variable. It can be easily shown that, if $X$ is exponentially distributed ($X \sim Exp(\lambda)$), then the first two moments of $X$ are $E[X] = \frac{1}{\lambda}$ and $E[X^2] = \frac{2}{\lambda^2}$. If $X$ is Poisson-distributed ($X \sim Poiss(\mu)$), then $E[X] = \mu$ and $E[X^2] = \mu + \mu^2$. By definition of the coalescent the $\xi_{kl}$ are distributed like $\xi_{kl} \sim Poiss(\frac{\theta}{2}T_k)$ with $T_k \sim Exp(\frac{2}{k(k-1)})$. $\xi_{kl}$ and $\xi_{k'l'}$ are independent if $k \neq k'$ while $\xi_{kl}$ and $\xi_{kl'}$ are independent conditional on $T_k$ for $l \neq l'$. We have thus

$$T_k \sim Exp(\lambda_k) \text{ with } \lambda_k = \frac{k(k-1)}{2} \text{ and } \xi_{kl} \sim Poiss(\mu) \text{ with } \mu = \frac{\theta}{2}T_k$$

$$E[\xi_{kl}] = E[E[\xi_{kl}|T_k]] = E[\frac{\theta}{2}T_k] = \frac{1}{k(k-1)}\theta$$

$$\begin{aligned}
E[\xi_{kl}^2] &= E[E[\xi_{kl}^2|T_k]] \\
&= E[\frac{\theta}{2}T_k + (\frac{\theta}{2}T_k)^2] \\
&= \frac{\theta}{2}E[T_k] + \frac{\theta^2}{4}E[T_k^2] \\
&= \frac{2}{k(k-1)}\frac{\theta}{2} + 2\frac{2^2}{k^2(k-1)^2}\frac{\theta^2}{4} \\
&= \frac{1}{k(k-1)}\theta + \frac{2}{k^2(k-1)^2}\theta^2
\end{aligned}$$

$$\begin{aligned}
E[\xi_{kl}\xi_{kl'}] =& E[E[\xi_{kl}\xi_{kl'}|T_k]] \\
&= E[E[\xi_{kl}|T_k]E[\xi_{kl'}|T_k]] \\
&= E[(\frac{\theta}{2}T_k)^2] \\
&= \frac{2}{k^2(k-1)^2}\theta^2
\end{aligned}$$

$$\begin{aligned}
E[\xi_{kl}\xi_{k'l}] &= E[\xi_{kl}]E[\xi_{k'l}] \\
&= \frac{1}{k(k-1)k'(k'-1)}\theta^2
\end{aligned}$$

For a particular topology, the number of mutations of size $i$ can be parcelled onto lines as

$$\xi_i = \sum_{k=2}^{n}\sum_{l=1}^{k}\epsilon_{kl}(i)\xi_{kl}$$

with the “indicator-variable” $\epsilon_{kl}(i) = 1$ if line $\xi_{kl}$ has $i$ descendent leaves and

0 otherwise. We take the expectation over all topologies and branch lengths:

$$\begin{aligned}
E[\xi_i \xi_j] =& E[(\sum_{k=2}^{n}\sum_{l=1}^{k} \epsilon_{kl}(i)\xi_{kl})(\sum_{k'=2}^{n}\sum_{l'=1}^{k'} \epsilon_{k'l'}(j)\xi_{k'l'})] \\
=& \sum_{k=2}^{n}\sum_{k'=2}^{n}\sum_{l=1}^{k}\sum_{l'=1}^{k'} E[\epsilon_{kl}(i)\epsilon_{k'l'}(j)]E[\xi_{kl}\xi_{k'l'}] \\
=& \sum_{k=2}^{n}\sum_{l=1}^{k} E[\epsilon_{kl}(i)\epsilon_{kl}(j)]E[\xi_{kl}\xi_{kl}] + \sum_{k=2}^{n}\sum_{l=1}^{n-1}\sum_{l'=l+1}^{n} E[\epsilon_{kl}(i)\epsilon_{kl'}(j)]E[\xi_{kl}\xi_{kl'}] + \\
& \sum_{k=2}^{n-1}\sum_{k'=k+1}^{n}\sum_{l=1}^{k}\sum_{l'=1}^{k'} \left(E[\epsilon_{kl}(i)\epsilon_{k'l'}(j) + E[\epsilon_{k'l}(i)\epsilon_{kl'}(j)]\right) E[\xi_{kl}\xi_{k'l'}] \\
=& \delta_{i=j}\sum_{k=2}^{n} kp(k,i)E[\xi_{kl}^2] + \sum_{k=2} k(k-1)p(k,i;k,j)E[\xi_{k1}\xi_{k2}] \\
& + \sum_{k=2}^{n-1}\sum_{k'=k+1}^{n} kk'(p(k,i;k',j) + p(k,j;k',i))E[\xi_{k1}\xi_{k'1}]
\end{aligned}$$

If we define for $k < k'$

$$\begin{aligned}
s_1(i) &= \sum_{k=2}^{n} kp(k,i)\frac{1}{k(k-1)} \\
s_2(i) &= \sum_{k=2}^{n} kp(k,i)\frac{2}{k^2(k-1)^2} \\
s(i,j) &= \sum_{k=2}^{n} k(k-1)p(k,i;k,j)\frac{2}{k^2(k-1)^2} \\
s_a(i,j) &= \sum_{k=2}^{n-1}\sum_{k'=k+1}^{n} kk'p_a(k,i;k',j)\frac{1}{k(k-1)k'(k'-1)} \\
s_b(i,j) &= \sum_{k=2}^{n-1}\sum_{k'=k+1}^{n} kk'p_b(k,i;k',j)\frac{1}{k(k-1)k'(k'-1)}
\end{aligned}$$

then

$$\begin{aligned}
E[\xi_i\xi_j] = \delta_{i=j}s_1(i)\,\theta + & \\
& (\delta_{i=j}s_2(i) + s(i,j) + s_a(i,j) + s_a(j,i) + s_b(i,j) + s_b(j,i))\,\theta^2\ .
\end{aligned}$$

The different relations between lines correspond to our subdivision of the conditional frequency spectrum. In particular, we have

$$\begin{aligned}
\mathrm{E}[\xi^{(sn)}_{i|j}] =& \delta_{i<j}\theta^2 j\, s_a(j,i)\\
\mathrm{E}[\xi^{(co)}_{i|j}] =& \delta_{i=j}\theta^2 j\, \left(s_2(i) + 2s_a(i,i)\right)\\
\mathrm{E}[\xi^{(en)}_{i|j}] =& \delta_{i>j}\theta^2 j\, s_a(i,j)\\
\mathrm{E}[\xi^{(cm)}_{i|j}] =& \delta_{i+j=n}\theta^2 j\, \left(s(i,j) + s_b(i,j) + s_b(j,i)\right)\\
\mathrm{E}[\xi^{(sd)}_{i|j}] =& \delta_{i+j<n}\theta^2 j\, \left(s(i,j) + s_b(i,j) + s_b(j,i)\right).
\end{aligned}$$

The following derivations simplify these expressions until we finally yield the equations 13.

The simplification makes use of two known formulas for binomial coefficients:

$$\sum_{m=0}^{n} \binom{m}{k} = \binom{n+1}{k+1} \tag{B1}$$

$$\sum_{j=0}^{k} \binom{m}{j}\binom{n-m}{k-j} = \binom{n}{k} \tag{B2}$$

In the first equation, the summation can start as well at $m = k$ since $\binom{m}{k} = 0$ for $m < k$.

Furthermore we need three helping equations from Fu (1995):

The straight-forward computable equation (14)

$$\frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} = \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}}\frac{k-1}{i}$$

and the technically more demanding equations (34)

$$2\sum_{k=2}^{n} \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}i}\frac{1}{k} = \beta_n(i)$$

and (36)

$$\sum_{k=3}^{n} \frac{\binom{n-i-2}{k-3}}{\binom{n-1}{k-1}} \frac{1}{k(k-1)} = \frac{\beta_n(i) - \beta_n(i+1)}{2} \quad .$$

A useful variation of equation (34), needed repeatedly, can be derived using his equation (33) (not replicated here):

$$\begin{aligned} \frac{1}{\binom{n-1}{i} i} \sum_{k=2}^{n} \frac{\binom{n-k}{i-1}}{k-1} &= \frac{1}{\binom{n-1}{i} i} \sum_{k=1}^{n-1} \frac{\binom{n-1-k}{i-1}}{k} \\ &\overset{(33)}{=} \frac{1}{n-i} \frac{1}{\binom{n-1}{i-1}} \binom{n-1}{i-1} (a_n - a_i) \\ &= \frac{a_n - a_i}{n-i} \quad . \end{aligned} \tag{34a}$$

Now we have to account for the “boundary cases” in the probability expressions $p()$. As defined above, a binomial coefficient $\binom{a}{b}$ with $a = -1$ is non-zero only for $b = -1$, which translates to additional constraints on the state $k$ and the number of descendants that the line from this state can have at state $k'$:

For example, if in the expression

$$p(k, i; k, j) = \frac{\binom{n-i-j-1}{k-3}}{\binom{n-1}{k-1}}$$

we have $i + j = n$, then the descendants of two lines encompass the whole sample. However this is only possible for the two lines of state $k = 2$. The same reasoning applied on $p_a(k, i; k', j)$ for $i = j$ leads to the condition, that the summation is only over one element, namely $t = 1$. Finally, if $i + j = n$ in the expression for $p_b(k, i; k', j)$, then $k = 2$ and $t = k' - 1$.

$$\begin{aligned}
s_1(i) &= \sum_{k=2}^{n} kp(k,i)\frac{1}{k(k-1)} \\
&\overset{(14)}{=} \sum_{k=2}^{n} k\frac{\binom{n-k}{i-1}(k-1)}{\binom{n-1}{i}i}\frac{1}{k(k-1)} \\
&= \frac{1}{i}\sum_{k=0}^{n-2}\frac{\binom{k}{i-1}}{\binom{n-1}{i}} \\
&\overset{(B1)}{=} \frac{1}{i}
\end{aligned}$$

$$\begin{aligned}
s_2(i) &= \sum_{k=2}^{n} kp(k,i)\frac{2}{k^2(k-1)^2} \\
&\overset{(14)}{=} \sum_{k=2}^{n} k\frac{\binom{n-k}{i-1}(k-1)}{\binom{n-1}{i}i}\frac{2}{k^2(k-1)^2} \\
&= \sum_{k=2}^{n}\frac{\binom{n-k}{i-1}}{\binom{n-1}{i}i}\frac{2}{k(k-1)} \\
&= 2\sum_{k=2}^{n}\frac{\binom{n-k}{i-1}}{\binom{n-1}{i}i}\left(\frac{1}{k-1}-\frac{1}{k}\right) \\
&\overset{(34),(34a)}{=} 2\frac{a_n-a_i}{n-i}-\beta_n(i)
\end{aligned}$$

$$\begin{aligned}
s(i,j) &= \sum_{k=2}^{n} k(k-1)p(k,i;k,j)\frac{2}{k^2(k-1)^2} \\
&= \sum_{k=2}^{n}\frac{\binom{n-i-j-1}{k-3}}{\binom{n-1}{k-1}}\frac{2}{k(k-1)} \\
&= \delta_{i+j<n}\sum_{k=3}^{n}\frac{\binom{n-(i-j-1)-2}{k-3}}{\binom{n-1}{k-1}}\frac{2}{k(k-1)}+\delta_{i+j=n}\frac{1}{n-1} \\
&\overset{(36)}{=} \delta_{i+j<n}\left(\beta_n(i+j-1)-\beta_n(i+j)\right)+\delta_{i+j=n}\frac{1}{n-1}
\end{aligned}$$

Case $i > j$ $(\Rightarrow t \geq 2)$

$$\begin{aligned}
s_a(i,j) &= \sum_{k'=3}^{n} \sum_{k=2}^{k'-1} kk' p_a(k,i;k',j) \frac{1}{k(k-1)k'(k'-1)} \\
&= \sum_{k'=3}^{n} \sum_{k=2}^{k'-1} \sum_{t=2}^{k'-1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{t}{k'} \frac{\binom{i-j-1}{t-2}\binom{n-i-1}{k'-t-1}}{\binom{n-1}{k'-1}} \frac{1}{(k-1)(k'-1)} \\
&\overset{(14)}{=} \sum_{k'=3}^{n} \sum_{k=2}^{k'-1} \sum_{t=2}^{k'-1} \frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}} \frac{\binom{i-j-1}{t-2}\binom{n-i-1}{k'-t-1}}{\binom{n-1}{k'-1}} \frac{1}{k'(k'-1)} \\
&= \sum_{k'=3}^{n} \sum_{t=2}^{k'-1} \frac{\binom{i-j-1}{t-2}\binom{n-j-2-(i-j-1)}{k'-3-(t-2)}}{\binom{n-1}{k'-1}} \frac{1}{k'(k'-1)} \sum_{k=1}^{k'-2} \frac{\binom{k}{t-1}}{\binom{k'-1}{t}} \\
&\overset{(B1)}{=} \sum_{k'=3}^{n} \sum_{t=0}^{k'-3} \frac{\binom{i-j-1}{t}\binom{n-j-2-(i-j-1)}{k'-3-t}}{\binom{n-1}{k'-1}} \frac{1}{k'(k'-1)} \\
&\overset{(B2)}{=} \sum_{k'=3}^{n} \frac{\binom{n-j-2}{k'-3}}{\binom{n-1}{k'-1}} \frac{1}{k'(k'-1)} \\
&\overset{(36)}{=} \frac{\beta_n(j) - \beta_n(j+1)}{2}
\end{aligned}$$

Case $i = j$ $(\Rightarrow t = 1)$

$$\begin{aligned}
s_a(i,i) &= \sum_{k'=3}^{n} \sum_{k=2}^{k'-1} kk' p_a(k,i;k',i) \frac{1}{k(k-1)k'(k'-1)} \\
&= \sum_{k'=3}^{n} \sum_{k=2}^{k'-1} \frac{\binom{k'-2}{k-2}}{\binom{k'-1}{k-1}} \frac{1}{k'} \frac{\binom{n-i-1}{k'-2}}{\binom{n-1}{k'-1}} \frac{1}{(k-1)(k'-1)} \\
&= \sum_{k'=3}^{n} \sum_{k=2}^{k'-1} \frac{k-1}{k'-1} \frac{1}{k'} \frac{\binom{n-i-1}{k'-2}}{\binom{n-1}{k'-1}} \frac{1}{(k-1)(k'-1)} \\
&= \sum_{k'=3}^{n} \frac{\binom{n-i-1}{k'-2}}{\binom{n-1}{k'-1}} \frac{k'-2}{k'(k'-1)^2} \\
&\overset{(14)}{=} \sum_{k'=3}^{n} \frac{\binom{n-k'}{i-1}}{\binom{n-1}{i}i} \frac{k'-2}{k'(k'-1)} \\
&= \sum_{k'=2}^{n} \frac{\binom{n-k'}{i-1}}{\binom{n-1}{i}i} \left(\frac{2}{k'} - \frac{1}{k'-1}\right) \\
&\overset{(34),(34a)}{=} \beta_n(i) - \frac{a_n - a_i}{n-i}
\end{aligned}$$

Case $i+j<n$ $(\Rightarrow t \leq k'-2)$

$$\begin{aligned}
s_b(i,j) &= \sum_{k'=3}^{n}\sum_{k=2}^{k'-1} kk'p_b(k,i;k',j)\frac{1}{k(k-1)k'(k'-1)} \\
&= \sum_{k'=3}^{n}\sum_{k=2}^{k'-1}\sum_{t=1}^{k'-2} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}}\frac{k'-t}{k'}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}}\frac{1}{(k-1)(k'-1)} \\
&\overset{(14)}{=} \sum_{k'=3}^{n}\sum_{k=2}^{k'-1}\sum_{t=1}^{k'-2} \frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}}\frac{k'-t}{tk'}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}}\frac{1}{k'-1} \\
&= \sum_{k'=3}^{n}\sum_{k=2}^{k'-1}\left(\sum_{t=2}^{k'-2}\frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}}\frac{k'-t}{tk'(k'-1)}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}} + \frac{1}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}}\right) \\
&= \sum_{k'=3}^{n}\left(\sum_{t=2}^{k'-2}\frac{k'-t}{tk'(k'-1)}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}}\sum_{k=1}^{k'-2}\frac{\binom{k}{t-1}}{\binom{k'-1}{t}} + \sum_{k=2}^{k'-1}\frac{1}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}}\right) \\
&\overset{(B1)}{=} \sum_{k'=3}^{n}\left(\sum_{t=2}^{k'-2}\frac{k'-t}{tk'(k'-1)}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}} + \frac{k'-2}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}}\right) \\
&= \sum_{k'=3}^{n}\sum_{t=2}^{k'-2}\frac{1}{t(k'-1)}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}} - \sum_{k'=3}^{n}\sum_{t=2}^{k'-2}\frac{1}{k'(k'-1)}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}} \\
&\quad + \sum_{k'=3}^{n}\frac{1}{k'-1}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}} - 2\sum_{k'=3}^{n}\frac{1}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}} \\
&= \sum_{k'=3}^{n}\sum_{t=1}^{k'-2}\frac{1}{t(k'-1)}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}} - \sum_{k'=3}^{n}\sum_{t=1}^{k'-2}\frac{1}{k'(k'-1)}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}} \\
&\quad - \sum_{k'=3}^{n}\frac{1}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}} \\
&= \frac{1}{i}\sum_{k'=3}^{n}\sum_{t=1}^{k'-2}\frac{1}{k'-1}\frac{\binom{i}{t}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}} - \sum_{k'=3}^{n}\sum_{t=1}^{k'-2}\frac{1}{k'(k'-1)}\frac{\binom{i-1}{t-1}\binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}} \\
&\quad - \sum_{k'=3}^{n}\frac{1}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}} \\
&= \frac{1}{i}\sum_{k'=3}^{n}\left(\sum_{t=0}^{k'-2}\frac{1}{k'-1}\frac{\binom{i}{t}\binom{n-j-1-i}{k'-2-t}}{\binom{n-1}{k'-1}} - \frac{1}{k'-1}\frac{\binom{n-i-j-1}{k'-2}}{\binom{n-1}{k'-1}}\right) \\
&\quad - \sum_{k'=3}^{n}\sum_{t=0}^{k'-3}\frac{1}{k'(k'-1)}\frac{\binom{i-1}{t}\binom{n-j-2-(i-1)}{k'-3-t}}{\binom{n-1}{k'-1}} - \sum_{k'=3}^{n}\frac{1}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}}
\end{aligned}$$

$$\begin{aligned}
&\overset{(B2)}{=}\frac{1}{i}\sum_{k'=3}^{n}\left(\frac{1}{k'-1}\frac{\binom{n-j-1}{k'-2}}{\binom{n-1}{k'-1}}-\frac{1}{k'-1}\frac{\binom{n-i-j-1}{k'-2}}{\binom{n-1}{k'-1}}\right)\\
&\quad-\left(\sum_{k=3}^{n}\frac{1}{k'(k'-1)}\frac{\binom{n-j-2}{k'-3}}{\binom{n-1}{k'-1}}+\sum_{k'=3}^{n}\frac{1}{k'(k'-1)}\frac{\binom{n-i-j-1}{k'-3}}{\binom{n-1}{k'-1}}\right)\\
&\overset{(14),(36)}{=}\frac{1}{i}\sum_{k'=2}^{n}\left(\frac{\binom{n-k'}{j-1}}{\binom{n-1}{j}}\frac{1}{j}-\frac{\binom{n-k'}{i+j-1}}{\binom{n-1}{i+j}}\frac{1}{i+j}\right)\\
&\quad-\frac{1}{2}\left(\beta_n(j)-\beta_n(j+1)+\beta_n(i+j-1)-\beta_n(i+j)\right)\\
&\overset{(B1)}{=}\frac{1}{ij}-\frac{1}{i(i+j)}-\frac{1}{2}\left(\beta_n(j)-\beta_n(j+1)+\beta_n(i+j-1)-\beta_n(i+j)\right)
\end{aligned}$$

Case $i+j=n$ ($\Rightarrow k=2$ and $t=k'-1$)

$$\begin{aligned}
s_b(n-j,j)&=\sum_{k'=3}^{n}k'p_b(2,n-j;k',j)\frac{1}{k'(k'-1)}\\
&=\sum_{k'=3}^{n}\frac{1}{k'-1}\frac{1}{k'}\frac{\binom{n-j-1}{t-1}}{\binom{n-1}{k'-1}}\frac{1}{k'-1}\\
&=\sum_{k'=3}^{n}\frac{\binom{n-j-1}{t-1}}{\binom{n-1}{k'-1}}\frac{1}{k'(k'-1)^2}\\
&\overset{(14)}{=}\sum_{k'=3}^{n}\frac{\binom{n-k'}{j-1}}{\binom{n-1}{j}j}\frac{1}{k'(k'-1)}\\
&=\sum_{k'=2}^{n}\frac{\binom{n-k'}{j-1}}{\binom{n-1}{j}j}\frac{1}{k'(k'-1)}-\frac{\binom{n-2}{j-1}}{\binom{n-1}{j}}\frac{1}{2j}\\
&=\sum_{k'=2}^{n}\frac{\binom{n-k'}{j-1}}{\binom{n-1}{j}j}\left(\frac{1}{k'-1}-\frac{1}{k'}\right)-\frac{1}{2(n-1)}\\
&\overset{(34a),(34)}{=}\frac{a_n-a_j}{n-j}-\frac{1}{2}\beta_n(j)-\frac{1}{2(n-1)}
\end{aligned}$$

## S.4 Derivation of the population spectrum

The population 1-SFS spectrum of linked sites $\mathrm{E}[\xi(f|f_0)]$ (equation 13) can be derived from the 1-SFS sample spectrum $\mathrm{E}[\xi_{k|l}]$ (equation 14) by the formula

$$\mathrm{E}[\xi(f|f_0)]=\lim_{n\to\infty}n\mathrm{E}[\xi_{\lfloor nf\rfloor|\lfloor nf_0\rfloor}]$$

The derivation is a cumbersome but relatively simple computation, once we prove a few limits and asymptotic results.

The "big O" notation $O(x_n)$ is used for a function of $n$ that behaves asymptotically as $x_n$ for $n \to \infty$, i.e. $O(x_n)/x_n \to$ constant. The indicator function $I(A)$ is 1 if $A$ is true and 0 otherwise.

First, we state two useful asymptotic results:

$$\lfloor nf \rfloor = nf \cdot (1 + O(1/n))$$

$$a_n = \ln(n) + \gamma + O(1/n) = (\ln(n) + \gamma) \cdot (1 + O(1/n \ln(n)))$$

where $\gamma$ is the Eulero-Mascheroni constant.

The main derivation involves the limits of a few terms. The first one is $nl(\beta_n(k) - \beta_n(k+1))$. By some manipulations:

$$\begin{aligned}
&nl(\beta_n(k) - \beta_n(k+1)) = \\
&= \frac{-4n^2 l(a_{n+1} - a_k)}{(n-k+1)(n-k)(n-k-1)} + \frac{2n^2 l(a_{k+1} - a_k)}{(n-k+1)(n-k)} + \frac{2nl}{(n-k)(n-k-1)} = \\
&= \left[ \frac{-4(l/n)(\ln(n+1) - \ln(k))}{(1-k/n+1/n)(1-k/n)(1-k/n-1/n)} + \frac{2(l/n)/(k/n)}{(1-k/n+1/n)(1-k/n)} + \right. \\
&\left. + \frac{2l/n}{(1-k/n)(1-k/n-1/n)} \right] \cdot (1 + O(1/n))
\end{aligned}$$

hence

$$n\lfloor nf_0 \rfloor (\beta_n(\lfloor nf \rfloor) - \beta_n(\lfloor nf \rfloor + 1)) \underset{n\to\infty}{\longrightarrow} \frac{4f_0 \ln(f)}{(1-f)^3} + \frac{2f_0/f}{(1-f)^2} + \frac{2f_0}{(1-f)^2}$$

The second one is $l\frac{a_n - a_k}{n-k}$:

$$l\frac{a_n - a_k}{n-k} = \frac{l}{n} \frac{\ln(n) - \ln(k)}{1 - k/n} \cdot (1 + O(1/n))$$

hence

$$\lfloor nf_0 \rfloor \frac{a_n - a_{\lfloor nf \rfloor}}{n - \lfloor nf \rfloor} \underset{n\to\infty}{\longrightarrow} f_0 \frac{-\ln(f)}{1-f}$$

The last one is $l\beta_n(k)$:

$$l\beta_n(k) = \frac{l}{n} \left( \frac{\ln(n) - \ln(k)}{(1-k/n)(1-k/n-1/n)} - \frac{2}{1-k/n} \right) \cdot (1 + O(1/n))$$

hence

$$\lfloor nf_0 \rfloor \beta_n(\lfloor nf \rfloor) \underset{n\to\infty}{\longrightarrow} f_0 \left( \frac{-2\ln(f)}{(1-f)^2} - \frac{2}{1-f} \right)$$

Finally, the limit of the Kronecker delta $\delta_{k,l}$ (which appears implicitly as a multiplicative factor in the spectrum for co-occurring and complementary mutations) is a non-trivial one. In fact, the limit

$$n\delta_{\lfloor nf \rfloor, \lfloor nf_0 \rfloor} \to \delta(f - f_0)$$

exists only as a convergence in the space of distributions. We prove it directly by showing that the two distributions converge when applied to an arbitrary smooth test function $h(f)$ with compact support:

$$\begin{aligned}
&\int_{-\infty}^{\infty} df\ h(f) \left[ n\delta_{\lfloor nf \rfloor, \lfloor nf_0 \rfloor} - \delta(f - f_0) \right] = \\
&= n \int_{-\infty}^{\infty} df\ h(f)\ I(\lfloor nf_0 \rfloor \leq nf < \lfloor nf_0 \rfloor + 1) - h(f_0) = \qquad \text{(use } x = nf - \lfloor nf_0 \rfloor\text{)} \\
&= \int_0^1 dx\ h(x/n + \lfloor nf_0 \rfloor / n) - h(f_0) \underset{n\to\infty}{\longrightarrow} h(f_0) - h(f_0) = 0
\end{aligned}$$

## S.5 Derivation of the triallelic spectrum

The mutation process for two non-recombining loci - resulting in the generation of three alleles - resembles the mutation process for a single triallelic locus once the different mutation rates are taken into account. The rescaled mutation rates for the two mutations are $(\theta_1^{loc}, \theta_2^{loc})$ instead of $(2\theta, \theta)$ for the two-site case (the first mutation can appear in either of the loci, hence the factor of 2). Moreover, we consider a single locus instead of $L(L-1)/2 \sim L^2/2$ pairs of sites. The overall factor is therefore $\theta_1^{loc}\theta_2^{loc}/\theta^2 L^2$. Both nested and disjoint components contribute to the triallelic spectrum.