The neutral frequency spectrum of linked sites

Luca Ferretti[a,b,*], Alexander Klassmann[c,*], Emanuele Raineri[d], Sebastián E. Ramos-Onsins[e], Thomas Wiehe[c], Guillaume Achaz[b]

[a] *The Pirbright Institute, Woking, United Kingdom*
[b] *Institut de Systématique, Evolution, Biodiversité, UMR 7205, MNHN and Centre Interdisciplinaire de Recherche en Biologie, UMR 7241, Collége de France.*
[c] *Institut für Genetik, Universität zu Köln, 50674 Köln, Germany*
[d] *CNAG-CRG, Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain*
[e] *Centre for Research in Agricultural Genomics (CRAG), Bellaterra, Barcelona, Spain*

Abstract

We introduce the conditional Site Frequency Spectrum (SFS) for a genomic region linked to a focal mutation of known frequency. An exact expression for its expected value is provided for the neutral model without recombination. Its relation with the expected SFS for two sites, 2-SFS, is discussed. These spectra derive from the coalescent approach of Fu (1995) for finite samples, which is reviewed. Remarkably simple expressions are obtained for the linked SFS of a large population, which are also solutions of the multi-allelic Kolmogorov equations. These formulae are the immediate extensions of the well known single site θ/f neutral SFS. Besides the general interest in these spectra, they relate to relevant biological cases, such as structural variants and introgressions. As an application, a recipe to adapt Tajima's D and other SFS-based neutrality tests to a non-recombining region containing a neutral marker is presented.

[*] These authors contributed equally
Email address: `luca.ferretti@gmail.com` (Luca Ferretti)

Keywords: site frequency spectrum, linkage disequilibrium, neutral evolution, Kingman coalescent

1. Introduction

One of the basic 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 $k \in [1, n-1]$. Sites with alleles at frequency 0 or 1 are not included in the SFS.

According to the standard neutral model of molecular evolution [1], 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 [2, 3]. Using coalescent theory, Fu [4] derived the mean and covariance matrix for each component 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 [5]. The SFS was also studied in scenarios including selection [6, 7], demography [8, 9] or population structure [10].

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 [11, 12, 13].

It was shown 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 [14]. 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. [15]). More sophisticated methods based on the expected SFS, such as Poisson Random Field [16, 17, 18] and Composite Likelihood approaches [*e.g.,* 7, 19, 20, 21], 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 knowledge on linked variants affects the distribution of summary statistics, therefore the spread (and possibly the mean) of the estimated parameters [22, 23]. For this reason, simulations of the evolution of linked sequences are required for an accurate estimation of the statistical support for different models [24].

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" [25], 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 [26, 27] as well as a numerical solution rely-

ing on simulations [25]. Even 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 locus) 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 non-recombinant 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 couplings between sites that is not available in the multi-allelic spectrum.

For an infinite population, the multi-alleles single-locus spectrum is the solution of a multiallelic diffusion equation [3, section 5.10]. Polynomial expansions were proposed to solve the diffusion equations for the SFS of an infinite population [28, 29, 30], as well as moment-based approaches [31]. Finally, a polynomial expansion of the 2-SFS has been found for two sites without recombination and with general selection coefficients [32]. 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 [33] 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 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 Jenkins, Song and collaborators [34, 35] for the Kingman coalescent and general allelic transition matrices. More recently, Sargsyan [36] generalized the result of [33] 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 review and present in its simplest possible form the exact solution for the expectation of the neutral sample 2-SFS without recombination, then we extend it to a closed-form solution for the continuous population 2-SFS. The solution for a finite sample was derived previously in many disguises in a coalescent framework [4, 34, 37, 36] and its extrapolation to the limit of infinite sample sizes yields the continuous spectrum, which is a solution of the multi-allelic Kolmogorov equations. Furthermore, we derive the expected 1-SFS of sites that are completely linked to a focal mutation of known frequency. This spectrum has several potential applications. In section S1 of the Supplementary Material we extend the formulae for the continuous 2-SFS to closed expressions for the multi-allelic spectrum of a locus with three alleles.

Finally, as an application, we present a recipe to build a class of SFS-based neutrality tests for sequences containing a known neutral marker of

given frequency. This is a typical scenario when the marker (and the region around it) has been detected independently as an outlier in genome-wide association studies or population differentiation studies with SNP arrays. As far as we know, this is the first proper adaptation of Tajima's D and similar statistical tests to this kind of sequence data.

Model definition and notation

We consider a population of N 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 μ 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 rates $\theta \ll 1$. More precisely, our results are derived in the limit $\theta \to 0$ with fixed non-zero θL, 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 ξ_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 ξ 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 fixed 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:

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

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:

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

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

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 $\mathrm{E}[\xi(f)] = \theta L/f$ [2, 3]. 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.

[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 $\Delta f \sim \frac{1}{N}$.

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 kind 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 multiple equivalent interpretations in the small θ 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. Here we always consider unordered pairs of sites (the ordered case is discussed in section S2).

The second kind of spectrum 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:

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

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

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 θ.

2. Results

2.1. Decomposition of the 2-SFS

We follow [36] 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 that are 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 the overall spectrum is not sufficient to provide a full description of the genetic state of the two sites, while the two components $\xi^{(n)}(f_1, f_2)$, $\xi^{(d)}(f_1, f_2)$ are enough to reconstruct the genetic

content of the two sites up to permutations of all the haplotypes, as it happens with the usual SFS for one site. For example, the following haplotypes (derived alleles marked in bold)

$$\begin{matrix} C\,T \\ C\,\boldsymbol{A} \\ \boldsymbol{G}\,\boldsymbol{A} \end{matrix} \quad \text{and} \quad \begin{matrix} C\,\boldsymbol{A} \\ C\,\boldsymbol{A} \\ \boldsymbol{G}\,T \end{matrix}$$

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:

- $\xi^{(sn)}(f|f_0)$: *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 carried by the same individuals;

[2]We subdivide the "strictly nested" mutations of [36] into *strictly nested* and *enclosing* mutations while we refer to "identical" mutations as *co-occurring*.

- $\xi^{(en)}(f|f_0)$: *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 exactly 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, as both types of branches are exclusive in a single tree.

With the above definition and using the rules of conditional probabilities $p(f, f_0) = p(f|f_0)p(f_0)$ and the interpretations discussed in the previous section, the relations between the two sets of population subspectra 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 present the conditional and joint spectra for the sample and the population. The derivations and proofs of all equations in this section

are given in Methods and sections S3 and **??** of the Supplementary Material. The folded version of the 2-SFS is provided in Appendix A.

2.2.1. The sample joint 2-SFS

The 2-loci spectrum appeared in the literature under many guises [4, 34, 37, 36]. In the infinite-sites neutral model without recombination, its expected value has a simpler form[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 [37] 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 combination 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 distribution concentrated in f_0 (i.e. $\delta(f - f_0) = 0$ for $f \neq f_0$, normalized such that $\int_{-\infty}^{\infty} \delta(f - f_0)df = 1$).

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:

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

Please note that hereafter unmet conditions imply 0 otherwise.

The strictly nested component of the conditional 1-SFS and its applications have been discussed by Griffiths and Tavare [38].

2.2.4. The population conditional 1-SFS

For the whole population, the expected linked SFS becomes:

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

2.3. Shape of the SFS

We report the full joint 2-SFS as well as the nested and disjoint components (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. Disjoint mutations are dominated by cases where both mutations are rare, 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 rare and their frequencies are uniformly distributed, as previously noted [33].

Finally, in Figure 4 we show the impact of a focal mutation of given frequency on two estimators of θ. The Watterson's estimator $\hat{\theta}_S$ [39] depends on the total number of polymorphic sites, which increases with the frequency of the focal mutation, as they inflate increasingly upper sections of the trees that contribute more to the total tree length. On the other hand, Tajima's estimator, $\hat{\theta}_\pi$ [40] is more sensitive to mutations of intermediate frequency. The difference between the two illustrates how the spectrum is skewed towards common or rare mutations. As Tajima's D [11] 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. Neutrality tests for regions linked to a polymorphic neutral marker

Biallelic putative neutral markers are often used to find regions of interest in a genome. For example, genotype data from SNP arrays can be used together with phenotype measurements to find regions associated with a specific phenotype. Alternatively, if data from multiple populations are available, markers with highly differentiated frequency between populations - i.e. high F_{st} - can be used to infer potential targets of local selection. It is then natural to use sequence data to test for neutral evolution in a window around the neutral focal marker of known frequency.

Up to now, such tests did not take into account the information given by the frequency of the marker itself. However, since typical markers have biased frequencies towards intermediate values [41], the expected neutral frequency spectrum will likely be dramatically altered. An example of the dependence of the distribution of Tajima's D on the frequency of the marker is shown in Figure S2.

The results of the previous section show precisely how Tajima's D test values for a fully linked locus is biased as a function of the frequency of the marker. These biases can be computed analytically for all Tajima's D-like SFS-based tests [13] in a similar way, using equation (14) and simple approximations (e.g. replacing S by its conditional expected value in their denominator).

Furthermore, it is possible to develop versions of Tajima's D and other frequency-spectrum based neutrality tests that take into account the presence of the neutral marker. The simplest approach follows Rafajlović et al. [42] and consists in replacing the neutral spectrum $\theta L/i$ by the conditional spectrum

$\xi_{i|m}$, where m is the count of the marker in the sample. The usual covariance of the spectrum from Fu [4], which appears in the normalisation of the tests, can be replaced by the one derived by Klassmann and Ferretti [43] for the conditional spectrum. We denote the neutral marker by ϕ and the allele count of its derived allele by m. We use the observed spectrum $\hat{\xi}_{i|\phi}$ for a window of size L containing the marker to build a test of the general form [13, 14]:

$$T_{\mathbf{\Omega}} = \frac{\sum_{i=1}^{n-1} \Omega_i \hat{\xi}_{i|\phi}/\xi_i^0}{\sqrt{\mathrm{Var}\left[\sum_{i=1}^{n-1} \Omega_i \hat{\xi}_{i|\phi}/\xi_i^0\right]}} \tag{15}$$

where both the null spectrum $\xi_i^0 = \mathrm{E}[\xi_{i|\phi}]/\theta$ and the variance in the denominator are computed under the standard neutral model, i.e. Kingman's coalescent. The real vector of parameters $\mathbf{\Omega}$ can be chosen in any possible way, as long as it satisfies $\sum_{i=1}^{n-1} \Omega_i = 0$. For example, in the absence of a neutral marker, Tajima's D corresponds to $\Omega_i = \frac{2(n-i)}{n(n-1)}$.

The definition of the test requires that the null spectrum and the variance are conditioned on the presence of the marker ϕ. For the neutral spectrum, it is simply the sum of the expected nested and disjoint spectra presented in equations 14:

$$\xi_i^0 = \mathrm{E}[\xi_{i|m}]/\theta = \frac{\mathrm{E}[\xi_{i|m}^{(n)}] + \mathrm{E}[\xi_{i|m}^{(d)}]}{\theta} \tag{16}$$

On the other hand, the variance can be decomposed as

$$\mathrm{Var}\left[\sum_{i=1}^{n-1} \Omega_i \hat{\xi}_{i|\phi}/\xi_i^0\right] = \sum_{i=1}^{n-1} \Omega_i^2 \theta/\xi_i^0 + \sum_{i,j=1}^{n-1} \Omega_i \Omega_j \mathrm{Cov}[\xi_{i|\phi}, \xi_{j|\phi}]_{\theta^2}/\xi_i^0 \xi_j^0 \tag{17}$$

where the θ contribution corresponds to the Poisson noise of the mutational process, while the θ^2 contribution to the covariance $\mathrm{Cov}[\xi_{i|\phi}, \xi_{j|\phi}]_{\theta^2}$ can be

easily obtained from the third moments of the spectrum derived by Klassmann and Ferretti [43]:

$$\begin{aligned}\mathrm{Cov}[\xi_{i|\phi}, \xi_{j|\phi}]_{\theta^2} = \mathrm{E}[\xi_{i|\phi}, \xi_{j|\phi}]_{\theta^2} - \theta^2 \xi_i^0 \xi_j^0 = \\ = \mathrm{E}[\xi_{i|\phi}^{(n)}, \xi_{j|\phi}^{(n)}]_{\theta^2} + \mathrm{E}[\xi_{i|\phi}^{(n)}, \xi_{j|\phi}^{(d)}]_{\theta^2} + \mathrm{E}[\xi_{i|\phi}^{(d)}, \xi_{j|\phi}^{(n)}]_{\theta^2} + \mathrm{E}[\xi_{i|\phi}^{(d)}, \xi_{j|\phi}^{(d)}]_{\theta^2} - \theta^2 \xi_i^0 \xi_j^0\end{aligned} \tag{18}$$

The test can then be built by putting together the results from the last section and an estimation of θ and θ^2. The Maximum Composite Likelihood estimate can be used:

$$\hat{\theta} = S / \sum_{i=1}^{n-1} \xi_i^0 \ , \ \hat{\theta^2} = (\hat{\theta})^2 \tag{19}$$

or the Method-of-Moments estimates as in the classical Tajima's D:

$$\hat{\theta} = S / \sum_{i=1}^{n-1} \xi_i^0 \ , \ \hat{\theta^2} = S(S-1) / \sum_{i,j=1}^{n-1} (\xi_i^0 \xi_j^0 + \mathrm{Cov}[\xi_{i|\phi}, \xi_{j|\phi}]_{\theta^2} / \theta^2) \tag{20}$$

The choice of weights for new tests of this form is someway arbitrary. For example, a modified version of Tajima's D could use the old weights, i.e. $\Omega_i = \frac{2(n-i)}{n(n-1)}$, or the old linear coefficients, i.e. $\Omega_i = \frac{2i(n-i)}{n(n-1)} \xi_i^0$, depending if the test should focus on the relative or absolute differences between the null and observed spectrum. Principles and formulae for a meaningful choice of the new weights are discussed in detail by Ferretti et al. [14]. On the other hand, once the weights are chosen, the normalisation does not suffer from any degree of arbitrariness and its form depends only on the third moments computed here.

This straightforward modification of neutrality tests is a promising direction for future dedicated neutrality tests that aim at correcting multiple artefacts such as demography, knowledge of the frequency of the marker, etc.

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

$$\mathrm{E}[\xi_{k,l}] = \mathrm{E}[\xi_k \xi_l], \text{ if } k \neq l \tag{21}$$

and

$$\mathrm{E}[\xi_{k,k}] = \mathrm{E}[\xi_k(\xi_k - 1)]/2. \tag{22}$$

These expected values have been derived by Fu [4] 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 [4] 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 [4]. This can be obtained directly by Fu's expression for the covariance matrix σ_{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 [4], tracking the parts that lead to our mutation classes, is provided in section S3 of the Supplementary Material.

The same results could also be obtained from Theorem 5.1 in Jenkins and Song [34]. In fact, for a special choice of allele transition matrices (in the tri-allelic case, a strictly lower triangular matrix with all non-zero entries equal to 1), their results for recurrent mutations for small θL (θ in their article)

are mathematically equivalent to the results for mutations in an infinite-sites model. 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 derivation makes use of 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 section S4.

4. Discussion

In this article, we have provided exact closed formulae for the joint 2-SFS as well as the first expressions for the conditional 1-SFS, both for sample and population. The 2-SFS was already derived in different forms [34, 37, 36], but the expression presented here for the infinite-sites model is embedded in the framework of Fu [4]. Sample spectra were then used 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 most existing mutations were "strictly disjoint". 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 highly unbalanced trees.

There are several potential applications of these results. Here we discuss approaches to correct SFS-based neutrality tests taking into account the presence of a neutral marker strongly linked to the genomic region. These corrections are useful in cases when the region has been selected on the basis of evidence from genome-wide association studies or studies of differentiation based on SNP arrays or other (putatively) neutral markers. This is just an example of possible extensions of neutrality tests based on these results.

Other applications include the improvement of population genetic inference techniques based on the SFS, such as composite likelihood [*e.g.* 7, 19, 20, 21] and Poisson Random Field methods [16]. 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, methods 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. As an example, it can be checked numerically that the expected value of D between fully linked derived mutations is 0 according to our equations, as expected from LD theory. 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. The spectra presented here could also provide a neutral model for other scenarios, including structural variants or introgressions from different species or populations. Introgressed sequences from close species can be detected as divergent haplotypes in the locus considered, and if introgressions are rare, then the genetic variability within these haplotypes is described by the nested spectrum linked to the introgressed haplotypes.

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 [9, 34] and to mutations in rapidly adapting populations using the Λ-coalescent approximation and the results of [44]. 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 [32] as a polynomial expansion, but the numerical computation of this expansion is still cumbersome. Given the simplicity of the expression for the single-locus SFS $\xi(f) = \theta(1 - e^{-2N_e s(1-f)})/f(1-f)(1-e^{-2N_e s})$ [2, 16], 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 [3, section 5.10]. In fact, 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 $1 - f_0$:

$$\frac{\partial \xi}{\partial t} = \frac{1}{2N_e}\left(\frac{\partial^2}{\partial f^2}\left[f(1-f)\xi\right] + 2\frac{\partial^2}{\partial f \partial f_0}\left[f(1-f_0)\xi\right] + \frac{\partial^2}{\partial f_0^2}\left[f_0(1-f_0)\xi\right]\right) \tag{23}$$

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 $1 - f_0 - f$:

$$\frac{\partial \xi}{\partial t} = \frac{1}{2N_e}\left(\frac{\partial^2}{\partial f^2}\left[f(1-f)\xi\right] - 2\frac{\partial^2}{\partial f \partial f_0}\left[f f_0 \xi\right] + \frac{\partial^2}{\partial f_0^2}\left[f_0(1-f_0)\xi\right]\right) \tag{24}$$

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. Our results could also be used to test the accuracy of existing tools based on a numerical solution of the diffusion equations [45].

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 [25], like expansions in the limit of strong recombination [46]. 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.

An immediate application of our results to recombination events is the following: since in the Ancestral Recombination Graph [47] 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.

We offer tools for computing the analytical spectra as well as performing simulations by manipulating output of the program *ms* [48]. The corresponding C++ code is contained in the package *coatli* developed by one of the authors and available on `http://sourceforge.net/projects/coatli/`.

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

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

[2] S. Wright, The distribution of gene frequencies under irreversible mutation, Proceedings of the National Academy of Sciences of the United States of America 24 (1938) 253.

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

[4] Y.-X. Fu, Statistical properties of segregating sites, Theoretical population biology 48 (1995) 172–197.

[5] L. Ferretti, A. Ledda, T. Wiehe, G. Achaz, S. E. Ramos-Onsins, Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests, Genetics (2017) genetics–116.

[6] J. C. Fay, C. I. Wu, Hitchhiking under positive darwinian selection, Genetics 155 (2000) 1405–13.

[7] Y. Kim, W. Stephan, Detecting a local signature of genetic hitchhiking along a recombining chromosome, Genetics 160 (2002) 765–77.

[8] R. C. Griffiths, S. Tavaré, Sampling theory for neutral alleles in a varying environment, Philos Trans R Soc Lond B Biol Sci 344 (1994) 403–10.

[9] D. Živković, T. Wiehe, Second-order moments of segregating sites under variable population size, Genetics 180 (2008) 341–357.

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

[11] F. Tajima, Statistical method for testing the neutral mutation hypothesis by dna polymorphism, Genetics 123 (1989) 585–95.

[12] Y. X. Fu, W. H. Li, Statistical tests of neutrality of mutations, Genetics 133 (1993) 693–709.

[13] G. Achaz, Frequency spectrum neutrality tests: one for all and all for one, Genetics 183 (2009) 249–58.

[14] L. Ferretti, M. Perez-Enciso, S. Ramos-Onsins, Optimal neutrality tests based on the frequency spectrum, Genetics 186 (2010) 353–65.

[15] X. Liu, Y.-X. Fu, Exploring population size changes using snp frequency spectra, Nature genetics 47 (2015) 555–559.

[16] S. A. Sawyer, D. L. Hartl, Population genetics of polymorphism and divergence., Genetics 132 (1992) 1161–1176.

[17] C. D. Bustamante, J. Wakeley, S. Sawyer, D. L. Hartl, Directional selection and the site-frequency spectrum, Genetics 159 (2001) 1779–88.

[18] C. D. Bustamante, R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, D. L. Hartl, The cost of inbreeding in arabidopsis, Nature 416 (2002) 531–4.

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

[20] Y. Kim, R. Nielsen, Linkage disequilibrium as a signature of selective sweeps, Genetics 167 (2004) 1513–24.

[21] R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, C. Bustamante, Genomic scans for selective sweeps using snp data, Genome Res 15 (2005) 1566–75.

[22] R. R. Hudson, et al., Gene genealogies and the coalescent process, Oxford surveys in evolutionary biology 7 (1990) 44.

[23] K. Thornton, Recombination and the properties of tajima's d in the context of approximate-likelihood calculation, Genetics 171 (2005) 2143–8.

[24] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional snp frequency data, PLoS Genet 5 (2009) e1000695.

[25] R. R. Hudson, Two-locus sampling distributions and their application, Genetics 159 (2001) 1805–1817.

[26] G. B. Golding, The sampling distribution of linkage disequilibrium, Genetics 108 (1984) 257–74.

[27] S. Ethier, R. Griffiths, On the two-locus sampling distribution, Journal of Mathematical Biology 29 (1990) 131–159.

[28] M. Kimura, Random genetic drift in a tri-allelic locus; exact solution with a continuous model, Biometrics 12 (1956) 57–66.

[29] R. Littler, E. Fackerell, Transition densities for neutral multi-allele diffusion models, Biometrics (1975) 117–123.

[30] R. Griffiths, A transition density expansion for a multi-allele diffusion model, Advances in Applied Probability (1979) 310–325.

[31] A. Hobolth, J. Siren, The multivariate wright–fisher process with mutation: Moment-based analysis and inference using a hierarchical beta model, Theoretical population biology 108 (2016) 36–50.

[32] X. Xie, The site-frequency spectrum of linked sites, Bulletin of mathematical biology 73 (2011) 459–494.

[33] A. Hobolth, C. Wiuf, The genealogy, site frequency spectrum and ages of two nested mutant alleles, Theoretical population biology 75 (2009) 260–265.

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

[35] P. A. Jenkins, J. W. Mueller, Y. S. Song, General triallelic frequency spectrum under demographic models with variable population size, Genetics 196 (2014) 295–311.

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

[37] L. Ferretti, E. Raineri, S. Ramos-Onsins, Neutrality tests for sequences with missing data, Genetics 191 (2012) 1397–1401.

[38] R. Griffiths, S. Tavare, The genealogy of a neutral mutation, Oxford Statistical Science Series (2003) 393–413.

[39] G. Watterson, On the number of segregating sites in genetical models without recombination, Theoretical population biology 7 (1975) 256–276.

[40] F. Tajima, Evolutionary relationship of dna sequences in finite populations, Genetics 105 (1983) 437–460.

[41] J. Lachance, S. A. Tishkoff, Snp ascertainment bias in population genetic analyses: why it is important, and how to correct it, Bioessays 35 (2013) 780–786.

[42] M. Rafajlović, A. Klassmann, A. Eriksson, T. Wiehe, B. Mehlig, Demography-adjusted tests of neutrality based on genome-wide snp data, Theoretical population biology 95 (2014) 1–12.

[43] A. Klassmann, L. Ferretti, The third moments of the site frequency spectrum, Theoretical population biology (2018).

[44] M. Birkner, J. Blath, B. Eldon, Statistical properties of the site-frequency spectrum associated with lambda-coalescents, Genetics (2013) genetics–113.

[45] A. P. Ragsdale, R. N. Gutenkunst, Inferring demographic history using two-locus statistics, Genetics 206 (2017) 1037–1048.

[46] P. A. Jenkins, Y. S. Song, Padé approximants and exact two-locus sampling distributions, The Annals of Applied Probability 22 (2012) 576–607.

[47] R. C. Griffiths, P. Marjoram, An ancestral recombination graph, Institute for Mathematics and its Applications 87 (1997) 257.

[48] R. R. Hudson, Generating samples under a wright-fisher neutral model of genetic variation, Bioinformatics 18 (2002) 337–8.

Appendix A. 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:

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

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:

$$\begin{aligned}
\mathrm{E}[\eta^{(n)}(f^*, f_0^*)] =& \mathrm{E}[\xi^{(n)}(f^*, f_0^*)] + \mathrm{E}[\xi^{(n)}(1-f^*, 1-f_0^*)] + \mathrm{E}[\xi^{(d)}(f^*, 1-f_0^*)] \\
&+ \mathrm{E}[\xi^{(d)}(1-f^*, f_0^*)] \\
\mathrm{E}[\eta^{(d)}(f^*, f_0^*)] =& \mathrm{E}[\xi^{(d)}(f^*, f_0^*)] + \mathrm{E}[\xi^{(n)}(f^*, 1-f_0^*)] + \mathrm{E}[\xi^{(n)}(1-f^*, f_0^*)]
\end{aligned} \tag{A.2}$$

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 co-occurring 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

$$\begin{aligned}
\mathrm{E}[\eta^{(sn)}(f^*|f_0^*)] =& \frac{f_0^*(1-f_0^*)}{\theta}\mathrm{E}[\eta^{(n)}(f^*, f_0^*)] \quad \text{for } f^* < f_0^* \\
\mathrm{E}[\eta^{(co)}(f^*|f_0^*)] =& 2 \cdot \frac{f_0^*(1-f_0^*)}{\theta}\mathrm{E}[\eta^{(n)}(f^*, f_0^*)] \quad \text{for } f^* = f_0^* \\
\mathrm{E}[\eta^{(en)}(f^*|f_0^*)] =& \frac{f_0^*(1-f_0^*)}{\theta}\mathrm{E}[\eta^{(n)}(f^*, f_0^*)] \quad \text{for } f^* > f_0^* \\
\mathrm{E}[\eta^{(cm)}(f^*|f_0^*)] =& 0 \\
\mathrm{E}[\eta^{(sd)}(f^*|f_0^*)] =& (1+\delta_{f^*,f_0^*}) \cdot \frac{f_0^*(1-f_0^*)}{\theta}\mathrm{E}[\eta^{(d)}(f^*, f_0^*)]
\end{aligned} \tag{A.3}$$

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:

$$\begin{aligned}
\mathrm{E}[\eta^{(n)}_{k^*,l^*}] =& \mathrm{E}[\xi^{(n)}_{k^*,l^*}] + \mathrm{E}[\xi^{(n)}_{n-k^*,n-l^*}] + \mathrm{E}[\xi^{(d)}_{k^*,n-l^*}] + \mathrm{E}[\xi^{(d)}_{n-k^*,l^*}] \\
\mathrm{E}[\eta^{(d)}_{k^*,l^*}] =& \mathrm{E}[\xi^{(d)}_{k^*,l^*}] + \mathrm{E}[\xi^{(n)}_{k^*,n-l^*}] + \mathrm{E}[\xi^{(n)}_{n-k^*,l^*}]
\end{aligned} \tag{A.4}$$

and the conditional 1-SFS is:

$$\begin{aligned}
\mathrm{E}[\eta^{(sn)}_{k^*|l^*}] =& \frac{l^*(n-l^*)}{\theta n}\mathrm{E}[\eta^{(n)}_{k^*,l^*}] \quad \text{for } k^* < l^* \\
\mathrm{E}[\eta^{(co)}_{k^*|l^*}] =& 2 \cdot \frac{l^*(n-l^*)}{\theta n}\mathrm{E}[\eta^{(n)}_{k^*,l^*}] \quad \text{for } k^* = l^* \\
\mathrm{E}[\eta^{(en)}_{k^*|l^*}] =& \frac{l^*(n-l^*)}{\theta n}\mathrm{E}[\eta^{(n)}_{k^*,l^*}] \quad \text{for } k^* > l^* \\
\mathrm{E}[\eta^{(cm)}_{k^*|l^*}] =& 0 \\
\mathrm{E}[\eta^{(sd)}_{k^*|l^*}] =& (1+\delta_{k^*,l^*}) \cdot \frac{l^*(n-l^*)}{\theta n}\mathrm{E}[\eta^{(d)}_{k^*,l^*}]
\end{aligned} \tag{A.5}$$

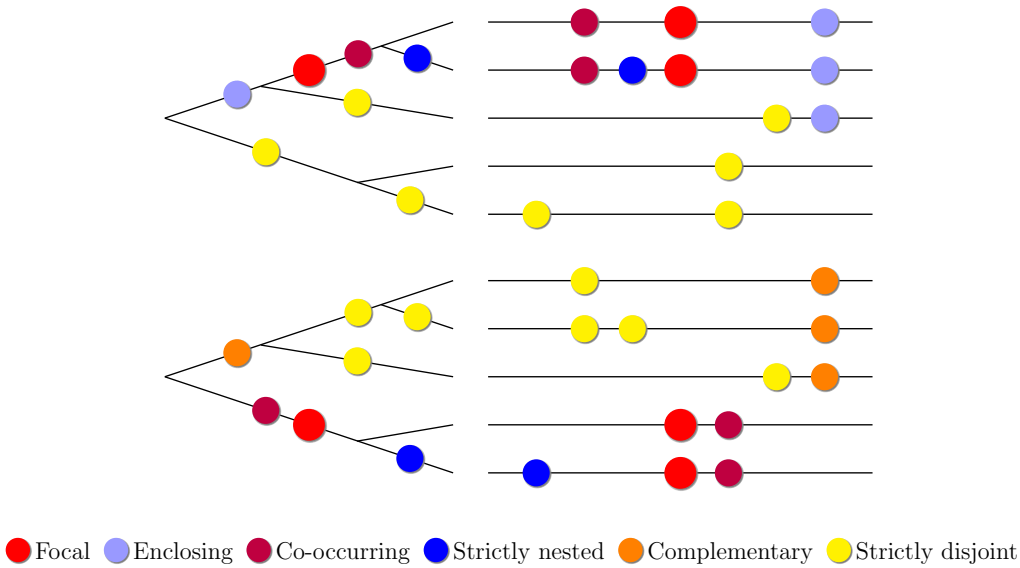

Figure 1: A schema of two non-recombining genomic regions and their corresponding genealogical trees. The black lines on the right represent sequences and the colored circles derived alleles. This figure illustrates the classification of all possible types of mutations with respect to the focal mutation (in red) and their occurrence on the sequence tree. If the focal mutation is not on a root branch (upper panel), it is clear 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 (lower panel), there cannot be enclosing mutations, but there can be mutations on the other root branch (*complementary*).

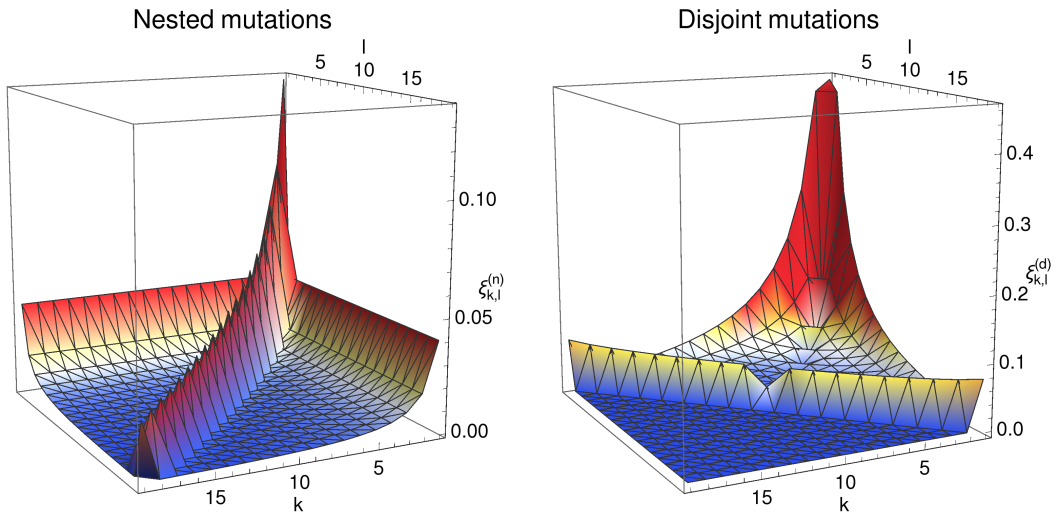

Figure 2: Plots of nested and disjoint contributions to the two-locus frequency spectrum for $\theta L = 1$, $n = 20$. Note the different scales of the two plots.

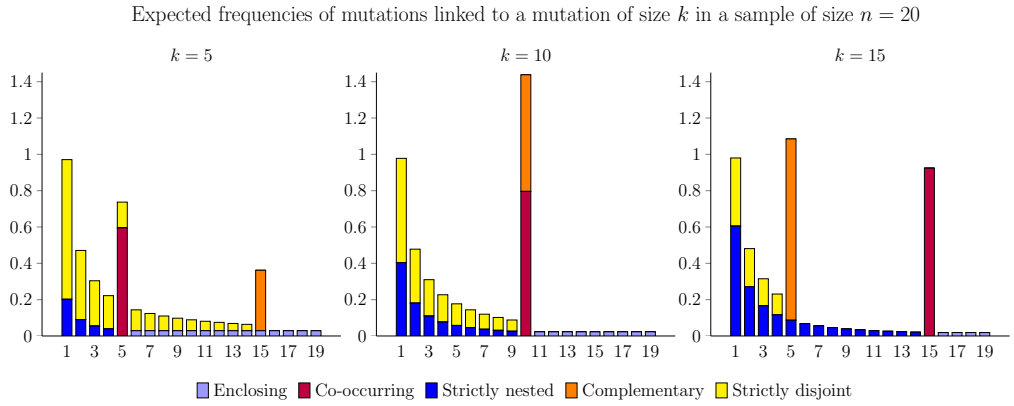

Figure 3: Barplot of the spectrum of linked sites for $\theta L = 1$, $n = 20$, each column colored according to the different contributions. The focal mutation has frequency 0.25 (left), 0.5 (middle) and 0.75 (right) respectively.

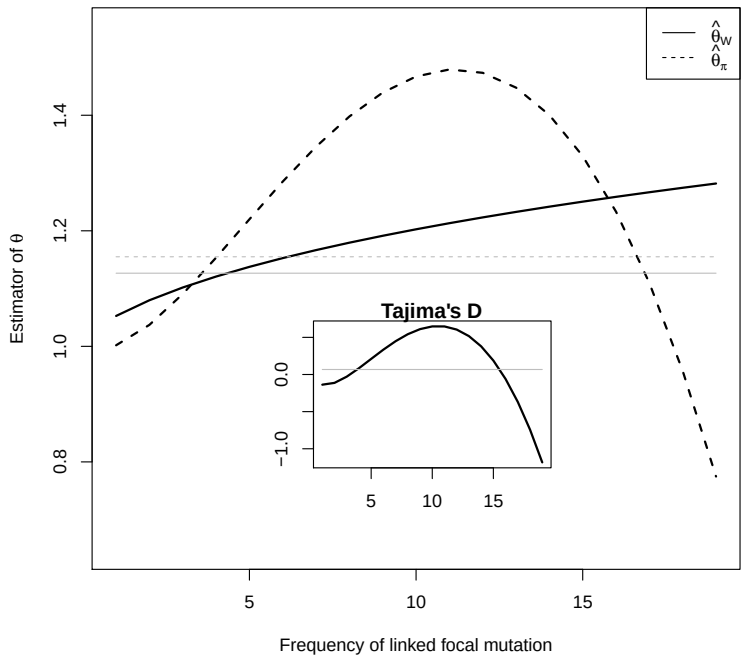

Figure 4: Mean values of the Watterson estimator ($\hat{\theta}_S$) and Tajima estimator ($\hat{\theta}_\pi$) of θ conditioned on the presence of a linked mutation, for $\theta = 1$, $n = 20$. In the inset, approximate mean value of Tajima's D (computed substituting S with its mean value in the denominator).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