Theoretical Population Biology 120 (2018) 16–28

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

# The third moments of the site frequency spectrum

A. Klassmann [a,*], L. Ferretti [b]

[a] Institut für Genetik, Universität zu Köln, 50674 Köln, Germany
[b] The Pirbright Institute, Woking, United Kingdom

ARTICLE INFO

*Article history:*
Received 20 February 2017
Available online 5 January 2018

*Keywords:*
Single nucleotide polymorphisms
Infinite-sites model
Site frequency spectrum
Coalescent approximation
Nested mutations
Skewness

ABSTRACT

The analysis of patterns of segregating (i.e. polymorphic) sites in aligned sequences is routine in population genetics. Quantities of interest include the total number of segregating sites and the number of sites with mutations of different frequencies, the so-called *site frequency spectrum*. For neutrally evolving sequences, some classical results are available, including the expected value and variance of the spectrum in the Kingman coalescent model without recombination as calculated by Fu (1995).

In this work, we use similar techniques to compute the third moments of the frequencies of three linked sites. Based on these results, we derive analytical results for the bias of Tajima's $D$ and other neutrality tests.

As a corollary, we obtain the second moments of the frequencies of two linked mutations conditional on the presence of a third mutation with a certain frequency. These moments can be used for the normalisation of new neutrality tests relying on these spectra.

## 1. Introduction

Statistics based on polymorphic loci are key to estimate relevant quantities in population genetics, such as the rescaled mutation rate $\theta$. One common approach is to group together variants that appear with the same frequency in a sample and count the elements of each such group. The resulting summary statistic is called the *site frequency spectrum*.

The frequency spectrum is one of the most relevant statistics for population genetics. It can be used to infer evolutionary parameters such as mutation and recombination rate, past population history, demography and selection (Hudson, 1983; Nielsen et al., 2005; Hein et al., 2004). Often, the variants are biallelic SNPs that can be "polarized", i.e. it is possible to say which allele is ancestral and which one is derived. This is the case for sequences with low mutation rate per base and for which an outgroup sequence is available. In what follows, we will consider exclusively this situation and assume that the evolution of these sequences can be modelled by a standard neutral Wright–Fisher model of constant population size.

Watterson (1975) credits Fisher (1930) with the first derivation (for a special case) of the first moments of the frequency spectrum. The derivation for the continuous analogue can be found in Ewens (1979), where it follows from results of diffusion theory (Kimura, 1964). Watterson (1975) himself derived the first and second moments for the sum over all classes of the frequency spectrum, i.e. the number of segregating sites, using the technique of "moment estimators". The full distribution of this quantity was shown by Tavaré (1984, Eq. (9.5)). The first and second moments for combinations of some components of the spectrum were later computed by Tajima (1989) using coalescent theory (Kingman, 1982) and combinatorics, while Fu (1995) completed this approach for the full frequency spectrum. A major application of his formulae is the normalisation of a class of neutrality tests such as Tajima's $D$ (Tajima, 1989), as described by Achaz (2009). Recently, Hudson (2015) has given another proof of the first moments. As far as we know, higher moments of the spectrum have never been computed.

Asymptotic results for the distribution of the spectrum have been obtained by Dahmer and Kersting (2015). However, their method applies only to mutations of size less than or equal to a fixed number $k$ in the limit of $n \to \infty$, i.e. to mutations of infinitesimal frequency $f \leq k/n \to 0$. Hence, their approach does not provide information on the full frequency spectrum in finite samples.

In this article we derive exact expressions for the third moments of the frequency spectrum. We use notation and approach of Fu (1995), with some technical modifications in order to keep the number of different cases manageable. As a by-product we state the third moment of the number of segregating sites. An immediate corollary of the third moments is the expected frequency spectrum for three linked segregating sites, which fully characterises the expected haplotype structure for triplets of sites.

We discuss the consequences of these results for the distribution of several neutrality tests that are constructed similarly to Tajima's $D$ (Tajima, 1989). These tests have been designed to

[*] Corresponding author.
*E-mail address:* alexander.klassmann@uni-koeln.de (A. Klassmann).

https://doi.org/10.1016/j.tpb.2017.12.002

yield under neutrality an expected value of approximately zero, but since they do not exactly so, they are biased (Tajima, 1989; Simonsen et al., 1995). For the first time, we obtain general expressions for bias and skewness of these tests as a function of mutation rate and sample size.

Finally, we derive the variance of the frequency spectra of two nested or disjoint mutations linked to a third mutation of a certain size. These spectra can be used to describe neutrally evolving structural variants such as chromosomal inversions (Ferretti et al., 2017). With our results, it is possible to obtain the proper normalisation for new Tajima's $D$-like tests relying on such spectra.

In the next section we state our main result and several implications. The corresponding proofs are presented largely in the subsequent section, while the combinatorial parts are deferred to the supplement.

## 2. Results

As is common practise in coalescent theory, we define $\theta$ as the population-scaled mutation rate per sequence, i.e. $\theta = 2pN_e\mu L$ where $p$ is the ploidy, $N_e$ is the effective population size, $\mu$ is the mutation rate per generation per bp and $L$ is the length of the sequence in base pairs. We consider a sample of $n$ sequences with $n \ll N_e$. We assume that we can distinguish between ancestral and derived alleles. A mutation (alias derived allele) is said to have size $i$, if $i$ sequences of the sample carry it. The number of mutations of size $i$ within the sample is referred to as $\xi_i$. The tuple $\xi_1, \ldots, \xi_{n-1}$ forms the frequency spectrum.

The model that we consider is the Kingman coalescent, with an infinite-sites model of mutations. We assume no recombination, i.e. complete linkage among sites.

### 2.1. The third moments of the frequency spectrum

Our main result is an analytical expression for the third moments of the frequency spectrum.

**Theorem 2.1.** *In the infinite sites approximation for biallelic sequences without recombination, the third moments of the frequency spectrum can be expressed as*

$$E[\xi_h\xi_i\xi_j] = \delta_{h=i=j}\tau_i\theta + \left(\delta_{h=i}\tau_{ij} + \delta_{i=j}\tau_{hj} + \delta_{j=h}\tau_{hi}\right)\theta^2 + \tau_{hij}\theta^3 \tag{1}$$

*for $1 \le h, i, j < n$. The functions $\tau$ are:*

$$\tau_i = \frac{1}{i}, \tag{2}$$

$$\tau_{ij} = t_a(i,j) + t_a(j,i) + t_b(i,j) + t_b(j,i) \tag{3}$$

*with*

$$t_a(i,j) = \begin{cases} \frac{1}{2}\left(\beta_n(j) - \beta_n(j+1)\right) & \text{if } j < i \\ \frac{1}{2}\beta_n(j) & \text{if } j = i \end{cases}$$
$$t_b(i,j) = \begin{cases} \frac{1}{ij} - \frac{1}{i(i+j)} - \frac{1}{2}\left(\beta_n(j) - \beta_n(j+1)\right) & \text{if } i+j < n \\ \alpha_n(j) - \frac{1}{2}\beta_n(j) & \text{if } i+j = n, \end{cases} \tag{4}$$

*and*[1]

$$\tau_{hij} = \sum_{\text{Permutations}(h,i,j)} t_{aa}(h,i,j) + t_{ab}(h,i,j) + t_{ba}(h,i,j) + t_{bb}(h,i,j) \tag{5}$$

*with Eqs. (6) given in Box I using the following auxiliary functions:*

$$\begin{aligned}
\alpha_n(i) &= \frac{1}{\binom{n-1}{i} i}\sum_{k=2}^{n}\frac{\binom{n-k}{i-1}}{k-1} \\
\beta_n(i) &= \frac{2}{\binom{n-1}{i} i}\sum_{k=2}^{n}\frac{\binom{n-k}{i-1}}{k} \\
\alpha_n^{(2)}(i,j) &= \sum_{k=2}^{n}\sum_{t=1}^{k-1}\frac{\binom{i-1}{t-1}\binom{n-i-j}{k-t-1}}{\binom{n-1}{k-1}}\frac{1}{k(k-1)}\alpha_k(t) \\
\beta_n^{(2)}(i,j) &= \sum_{k=2}^{n}\sum_{t=1}^{k-1}\frac{\binom{i-1}{t-1}\binom{n-i-j}{k-t-1}}{\binom{n-1}{k-1}}\frac{1}{k(k-1)}\frac{\beta_k(t)}{2} \\
\alpha_n^{(3)}(h,i,j) &= (h+1)\alpha_n^{(2)}(i,j) - 2h\alpha_n^{(2)}(i,j+1) \\
&\quad + (h-1)\alpha_n^{(2)}(i,j+2) \\
\beta_n^{(3)}(h,i,j) &= (h+1)\beta_n^{(2)}(i,j) - 2h\beta_n^{(2)}(i,j+1) \\
&\quad + (h-1)\beta_n^{(2)}(i,j+2) \\
\alpha_n^{(4)}(h,i,j) &= (h+1)\alpha_n^{(2)}(i+1,j) - 2h\alpha_n^{(2)}(i,j+1) \\
&\quad + (h-1)\alpha_n^{(2)}(i-1,j+2) \\
\beta_n^{(4)}(h,i,j) &= (h+1)\beta_n^{(2)}(i+1,j) - 2h\beta_n^{(2)}(i,j+1) \\
&\quad + (h-1)\beta_n^{(2)}(i-1,j+2).
\end{aligned} \tag{7}$$

**Remark 1.** The coefficient for $\theta$ is the well known result for the expectation of the frequency spectrum

$$E[\xi_i] = \tau_i\theta = \frac{\theta}{i}. \tag{8}$$

The terms $\tau_{ij}$ are identical to the quadratic part of the second moments,

$$E[\xi_i\xi_j] = \delta_{i=j}\tau_i\theta + \tau_{ij}\theta^2\,, \tag{9}$$

computed by Fu (1995): $\tau_{ij} = \sigma_{ij} + \frac{1}{ij}$, with $\sigma_{ij}$ defined in Eqs. (2) and (3) therein.

**Remark 2.** Fu (1995) showed in his Eq. (34) that $\alpha_n(i)$ and $\beta_n(i)$ can be written in a more compact form, namely

$$\begin{aligned}
\alpha_n(i) &= \frac{H_{n-1} - H_{i-1}}{n-i} \\
\beta_n(i) &= \frac{2n}{(n-i+1)(n-i)}(H_n - H_{i-1}) - \frac{2}{n-i},
\end{aligned}$$

with $H_n = \sum_{i=1}^{n}\frac{1}{i}$. We do not have a corresponding form for $\alpha_n^{(2)}(i,j)$ and $\beta_n^{(2)}(i,j)$. We only note that in the case of "singletons" they yield (with $H_{n,2} = \sum_{k=1}^{n}\frac{1}{k^2}$)

$$\begin{aligned}
\alpha_n^{(2)}(1,1) &= \frac{H_{n-1,2} - \frac{1}{n}H_{n-1}}{n-1} \\
\beta_n^{(2)}(1,1) &= \frac{1 - \frac{1}{n}H_{n-1}}{n-1}.
\end{aligned}$$

**Remark 3.** The sum over permutations simplifies the fractions in $t_b$ resp. $t_{bb}$:

$$\sum_{\text{Permutations}(i,j)}\left(\frac{1}{ij} - \frac{1}{i(i+j)}\right) = \frac{1}{ij} \tag{10}$$

[1] $\sum_{Perm.(h,i,j)} f(h,i,j) = f(h,i,j) + f(i,j,h) + f(j,h,i) + f(h,j,i) + f(i,h,j) + f(j,i,h)$.