The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele

Paul A. Jenkins[a,*], Yun S. Song[a,b]

[a] *Computer Science Division, University of California, Berkeley, Berkeley CA 94720, USA*

[b] *Department of Statistics, University of California, Berkeley, Berkeley CA 94720, USA*

Abstract

The sample frequency spectrum of a segregating site is the probability distribution of a sample of alleles from a genetic locus, conditional on observing the sample to be polymorphic. This distribution is widely used in population genetic inferences, including statistical tests of neutrality in which a skew in the observed frequency spectrum across independent sites is taken as a signature of departure from neutral evolution. Theoretical aspects of the frequency spectrum have been well studied and several interesting results are available, but they are usually under the assumption that a site has undergone at most one mutation event in the history of the sample. Here, we extend previous theoretical results by allowing for at most *two* mutation events per site, under a general finite alleles model in which the mutation rate is independent of current allelic state but the transition matrix is otherwise completely arbitrary. Our results apply to both nested and nonnested mutations. Only the former has been addressed previously, whereas here we show it is the latter that is more likely to be observed except for very small sample sizes. Further, for *any* mutation transition matrix, we obtain the joint sample frequency spectrum of the two mutant alleles at a triallelic site, and derive a closed-form formula for the expected age of the younger of the two mutations given their frequencies in the population. Several large-scale resequencing projects for various species are presently under way and the resulting data will include some triallelic polymorphisms. The theoretical results described in this paper should prove useful in population genomic analyses of such data.

*Corresponding author

Keywords: Frequency spectrum, Coalescent, Triallelic site, Genealogy, Allele age

1. Introduction

The frequency spectrum for a sample of genetic data taken from a population is a useful statistic, containing more information than single-value summaries like the number of segregating sites, yet remaining more tractable than working with the full data configuration. The sample frequency spectrum for a polymorphic site is defined as the probability distribution of the number of copies of the derived, or *mutant*, allele in a sample of size n. For a sample with many polymorphic sites, a histogram of the number of sites with i copies of the mutant allele present in the sample, for each $i = 1, \ldots, n-1$, can be compared to the sample frequency spectrum. In this manner, one can test the applicability of a given reproductive model by comparing departures of the observed frequency spectrum from its expectation. Most of the widely-used tests of neutrality are either directly or indirectly based on this observation (Achaz, 2009). Under a standard, neutral, coalescent model, and assuming the infinite sites model of mutation, the sample frequency spectrum is known in closed-form:

$$\phi(i) = \frac{i^{-1}}{\sum_{j=1}^{n-1} j^{-1}}, \tag{1}$$

where $\phi(i)$ is the probability that a mutant allele is present in exactly i copies of the sample [Watterson (1975); see Fu (1995), Griffiths and Tavaré (1998) for a coalescent approach]. This appealing result has been generalized to a number of further settings, including variable population size (Griffiths and Tavaré, 1998; Polanski and Kimmel, 2003; Evans et al., 2007), and genic selection (Griffiths, 2003). Bustamante et al. (2001) obtain a number of results related to the frequency spectrum for mutant sites under selection in the Poisson random field model of Sawyer and Hartl (1992).

All of this work assumes the infinite sites model of mutation. In particular, the mutation giving rise to the new allele is assumed to have occurred at most *once* in the genealogy relating the sample. Since the per-site mutation parameter θ is small (typically $0.001 \leq \theta \leq 0.01$ for humans, where $\theta = 4Nu$, N is the diploid effective population size, and u is the probability

of a mutation event per individual per generation), this assumption is usually reasonable. Occasionally, however, one might observe a site that must have undergone more than one mutation: it may be triallelic, or may be incompatible with the gene genealogy inferred from completely linked sites. Moreover, recurrent mutations can affect sites that still appear to conform to the infinite sites assumption. Thus far there have been no clear theoretical grounds for how to deal with nonconforming sites when working with the frequency spectrum. For example, one simple solution is simply to bin both the mutant alleles of a triallelic site and then to treat it as if it were diallelic (e.g. Johnson and Slatkin, 2006), but this is clearly not ideal.

In this work we obtain a more general distribution for the number of copies of mutant alleles at a site, by allowing at most *two* mutation events in the genealogy relating the sample. We employ a general *finite* sites model in which a fixed but arbitrary number of alleles, K, may be observed at the site of interest, and mutations between alleles occur according to some transition matrix P. The sample frequency spectrum is then more generally defined to be the joint probability distribution of the number of copies of each of the mutant alleles, conditional on at least one mutant allele. We assume the standard coalescent (Kingman, 1982), and derive our results by arguments using topological constraints induced on the genealogy by the two mutations. This approach is most closely related to the work of Wiuf and Donnelly (1999) and Hobolth and Wiuf (2009), who studied genealogies with one mutation and genealogies with two nested mutations, respectively. Among other results, Wiuf and Donnelly (1999) obtain the density of the age of a single mutant allele given its population frequency, and Hobolth and Wiuf (2009) obtain the joint and marginal sample frequency spectra of two mutant alleles when the mutations are genealogically nested, and the age of the younger of the two nested mutants. In this paper we extend these results to nonnested mutations, which, as we show below, is the more important of the two cases: With increasing sample size, the probability that two mutations are nonnested approaches one, and it is even the most probable outcome for all sample sizes greater than four. With results for both cases in hand, by averaging over whether or not the mutations are nested we obtain the sample frequency spectra of two mutant alleles regardless of their topological placement in the genealogy. Furthermore, Hobolth and Wiuf (2009) treat the two mutants as having occurred at two completely linked but distinct sites, so that the younger and older of the two mutants are always identifiable. In this work we model the two mutations as occurring at the

same site, allowing for the more general possibility of parallel mutations or back mutations. Particular choices of P in our model allows one to include the setting of Hobolth and Wiuf (2009) as a special case.

When introducing a model for mutation there are two cases to consider:

1. The allele of the most recent common ancestor (MRCA) of the sample is known, usually by comparison with an outgroup that is related by a suitable evolutionary distance.
2. The type of the MRCA is unknown.

In this work we largely restrict ourselves to the *first* case. In principle it has more power, since mutant alleles observed i times and $n - i$ times are distinguishable. When we have no prior assumptions regarding which of the alleles is the mutant, one must resort to the *folded* frequency spectrum, in which the two categories are binned together. In any case, when the type of the MRCA is unknown and the mutation transition matrix takes on a special *parent-independent* form—that is, P_{ij} is independent of i, for each pair of alleles i and j—then a closed-form sampling distribution for each site is available, which applies for any number of mutations in the history of the sample. This formula is essentially due to Wright (1949). Use of Wright's formula for making inferences regarding the site frequency spectrum is considered by Desai and Plotkin (2008). Note that when we assume the allele of the MRCA is known, Wright's formula does not apply even when mutation is parent-independent. For larger mutation rates, the assumption that a genealogy has undergone at most two mutations and that the allele of the MRCA is known each becomes less justifiable, and without prior information about which allele is mutant one should revert to using a folded site frequency spectrum.

In the special case of parent-independent mutation with the type of the MRCA unknown one can use Wright's formula as described above. It also applies to a diallelic model ($K = 2$), which can always be transformed into an equivalent parent-independent one. Aside from these cases, there are no classical results for the sample frequency spectrum under more general transition matrices. In this work we allow P to remain a general transition matrix apart from the restriction that the mutation rate at the locus is independent of its current allelic state. This is equivalent to ensuring $P_{ii} = 0$ for each $i = 1, \ldots, K$, since the effective rate at which an allele mutates to another distinct allele is $(\theta/2)(1 - P_{ii})$; in more general mutation models, P_{ii} can vary for different i to allow different rates of transition out of differ-

ent allelic states. It should be possible to modify our results to relax this assumption, albeit with a noticeable cost in bookkeeping, and so we do not attempt it in this work. An exception can be made when we study triallelic sample configurations later in the paper: genealogies associated with such configurations must have undergone at least two nontrivial mutation events, and we can allow $P_{ii} > 0$ without any additional effect. Essentially, having observed a triallelic sample together with the assumption that there were at most two mutation events means we condition on such trivial mutations not having occurred, even if we allow them back into the model. When studying triallelic configurations, we also find that our results simplify substantially with the following additional assumption:

$$P_{ab} = P_{cb}, \quad \text{and} \quad P_{ac} = P_{bc}, \tag{2}$$

where the three observed alleles are a, b, and c, and a is the ancestral allele. The condition (2) is satisfied by, and is weaker than, parent-independent mutation. It requires only that parent-independence holds in relevant entries of P, namely, in the rates of transition to each of the observed mutant alleles from the ancestral allele and from the other observed mutant allele.

Our paper is structured as follows. In Section 2 we introduce the recursion relation for the distribution of the sample configuration, which is well-known and is based on coalescent arguments. We utilize this recursion to obtain results for coalescent trees with one mutation event (Section 3) and two mutation events (Section 4). These results are made tractable in Section 5 by letting the mutation parameter go to zero, from which we can obtain useful expressions when we condition on certain observed patterns (e.g. that the site is triallelic) in Section 6. In Section 7 we also investigate the mean age in the population of a mutant allele at a triallelic site, and in Section 8 we investigate the accuracy of our expressions when the mutation parameter is in fact nonzero. We conclude with some brief discussion in Section 9.

2. Sample recursion

Denote an unordered sample configuration at a particular site by $\boldsymbol{n} = (n_1, n_2, \ldots, n_K)$, where K is the fixed and known number of alleles which could be observed at this site, and denote the sample size by $n = \sum_{i=1}^{K} n_i$. Members of the sample are referred to as *gametes*, so that n_i denotes the number of gametes in the sample with allele i. We fix the ancestral allele and

denote it as $a \in \{1, \ldots, K\}$. Denote by E_s the event that there are exactly s mutation events in the history of the sample. We will write the probability of observing the configuration $\boldsymbol{n}$ as $p(\boldsymbol{n})$, and the joint probability of this configuration together with E_s as $p(\boldsymbol{n}, E_s)$. It is implicit in these expressions that we condition on the ancestral allele being a. We can now obtain the sample frequency spectrum from these probabilities. For example, suppose we have two possible alleles, denoted 1 and 2 where 1 is ancestral, so that $P = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. As $\theta \to 0$, we may assume the mutant allele arose as the result of precisely one mutation event, and the distribution (1) is recovered as

$$\phi(i) = \lim_{\theta \to 0} p((n-i, i) \mid E_1) = \lim_{\theta \to 0} \frac{p((n-i, i), E_1)}{\sum_{j=1}^{n-1} p((n-j, j), E_1)},$$

$i = 1, \ldots, n-1$, since $p((n-i, i), E_1) = \theta i^{-1} + O(\theta^2)$ under this model (see (17) below).

Define a *history* to be the sequence of configurations $\boldsymbol{n} \mapsto \boldsymbol{n}' \mapsto \ldots \mapsto \boldsymbol{e}_a$ as we trace the ancestry of the sample back in time. Here, $\boldsymbol{e}_i$ denotes a sample comprised of a single gamete whose allele is i, so $\boldsymbol{e}_a$ denotes the sample comprising only the MRCA. A history can be regarded as an equivalence class in the space of genealogies with mutations that relate the sample. At each step in the sequence, at which the configuration is modified, we do not record *which* lineages are involved in each event, so there are many possible genealogies associated with any given history.

The probability $p(\boldsymbol{n})$ satisfies the following recursion relation for $n \geq 2$:

$$\begin{aligned} p(\boldsymbol{n}) = & \sum_{j=1}^{K} \frac{n_j - 1}{n - 1 + \theta} p(\boldsymbol{n} - \boldsymbol{e}_j) \\ & + \frac{\theta}{n - 1 + \theta} \sum_{i=1}^{K} \sum_{j=1}^{K} P_{ij} \frac{n_i + 1 - \delta_{ij}}{n} p(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i), \end{aligned} \tag{3}$$

with boundary condition $p(\boldsymbol{e}_i) = \delta_{ia}$ for each $i = 1, \ldots, K$, where δ_{ij} is the Kronecker delta. Similarly, the probability $p(\boldsymbol{n}, E_s)$ satisfies, for $n \geq 2$,

$$\begin{aligned} p(\boldsymbol{n}, E_s) = & \sum_{j=1}^{K} \frac{n_j - 1}{n - 1 + \theta} p(\boldsymbol{n} - \boldsymbol{e}_j, E_s) \\ & + \frac{\theta}{n - 1 + \theta} \sum_{i=1}^{K} \sum_{j=1}^{K} P_{ij} \frac{n_i + 1 - \delta_{ij}}{n} p(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i, E_{s-1}), \end{aligned} \tag{4}$$

with boundary conditions $p(\boldsymbol{e}_i, E_s) = \delta_{ia}\delta_{s0}$. Equations (3) and (4) are obtained from similar ones in Griffiths and Tavaré (1994), with slight modifications to the boundary conditions to reflect that we consider the allele of the MRCA to be known. Each path back through the recursions (3) and (4) is associated with a particular history, and we use this observation to obtain our results for the sample frequency spectrum. To illustrate the method, in Section 3 we first consider histories with precisely one mutation event. Henceforth we assume that $P_{ii} = 0$ for $i = 1, \ldots, K$, so that mutation events always result in a change of allele, and subsequent configurations in any history are always distinct. Finally, we also use the following notation: for a nonnegative real number x and a positive integer k,

$$(x)_k := x(x+1)\ldots(x+k-1)$$

denotes the kth ascending factorial of x.

3. One mutation event

Refer to the intervals back in time while there existed $n, n-1, \ldots, 2$ ancestors to the sample as *levels*. Wiuf and Donnelly (1999) proceed by conditioning on the level at which the unique mutation event occurred, and then considering the distribution of the number of offspring of each lineage from that level. Here, we take a related approach but instead argue directly from (4).

Suppose we observe the sample configuration $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b$, where $b \neq a$ is some mutant allele, and we have $n_a > 0$, $n_b > 0$, and $n_a + n_b = n$. Denote the sample configuration immediately before (more recently than) the time of the unique mutation event by $\boldsymbol{l}$. Conditional on the event E_1, that exactly one mutation event occurred in the history of the sample, $\boldsymbol{l}$ must be of the form $\boldsymbol{l} = l_a\boldsymbol{e}_a + \boldsymbol{e}_b$ for some l_a with $1 \leq l_a \leq n_a$ (Figure 1). We refer to a history that passes through the state $\boldsymbol{l}$ as *compatible* with $\boldsymbol{l}$. Thus, a history $\mathcal{H}_1$ that gives rise to $\boldsymbol{n}$, is compatible with $\boldsymbol{l}$, and is consistent with E_1, must be a sequence of configurations of the form $\boldsymbol{n} \mapsto \ldots \mapsto \boldsymbol{l} \mapsto (l_a+1)\boldsymbol{e}_a \mapsto \ldots \mapsto \boldsymbol{e}_a$. We make use of the following simple but useful lemma.

Lemma 3.1. *Conditional on E_1, on the observed sample configuration $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b$, and on the configuration $\boldsymbol{l} = l_a\boldsymbol{e}_a + \boldsymbol{e}_b$ at the time of the mutation event, the distribution of compatible histories $\mathcal{H}_1$ is* uniform. *It is given by*

$$p(\mathcal{H}_1 \mid E_1, \boldsymbol{n}, \boldsymbol{l}) = \binom{n - l_a - 1}{n_b - 1}^{-1}.$$

Moreover, the (unconditional) probability of such histories is

$$p(\mathcal{H}_1) = p(\mathcal{H}_1, E_1, \boldsymbol{n}, \boldsymbol{l}) = \theta P_{ab} \frac{(n_a - 1)!(n_b - 1)!}{(1+\theta)_{n-1}} \cdot \frac{l_a}{l_a + \theta}. \tag{5}$$

A proof of Lemma 3.1, along with our other results, is given in Appendix A. Summing over the $\binom{n-l_a-1}{n_b-1}$ histories and over $\boldsymbol{l}$, we obtain

$$\begin{aligned} p(\boldsymbol{n}, E_1) &= \sum_{l_a=1}^{n_a} \binom{n - l_a - 1}{n_b - 1} p(\mathcal{H}_1, E_1, \boldsymbol{n}, \boldsymbol{l}) \\ &= \begin{cases} \theta P_{ab} \dfrac{(n-1)!}{(1+\theta)_{n-1}} \displaystyle\sum_{l_a=1}^{n_a} \frac{\binom{n_a-1}{l_a-1}}{\binom{n-1}{l_a}} \cdot \frac{1}{l_a + \theta}, \\ \qquad \text{if } \boldsymbol{n} = n_a \boldsymbol{e}_a + n_b \boldsymbol{e}_b, \text{ where } b \neq a \text{ and } 1 \leq n_a, n_b \leq n, \\ 0, \quad \text{otherwise.} \end{cases} \end{aligned} \tag{6}$$

We now extend this approach to histories with precisely *two* mutation events.

4. Two mutation events

There are four cases to consider (Figure 2). The two mutation events are either nested (denoted $E_{2\mathcal{N}}$), nonnested ($E_{2\mathcal{NN}}$), on the same edge ($E_{2\mathcal{S}}$), or basal ($E_{2\mathcal{B}}$). We define each of these in further detail below; for now note that *nested* excludes the case that the mutations occurred on the same edge, and *nonnested* excludes the case that the mutations reside on the two basal (innermost) edges of the tree. We use superscript notation to further specify the alleles to which the two age-ordered mutation events gave rise, so for example $E_{2\mathcal{N}}^{(b,c)}$ ($\subseteq E_{2\mathcal{N}}$) denotes the event that there were precisely two mutation events, the mutations were nested, that the older mutation gave rise to a b allele, and that the younger mutation gave rise to a c allele. Note that in this example we must have $a \neq b$ and $b \neq c$ but it may or may not be the case that $a = c$. The $a = c$ case will be dealt with separately. Similar special cases arise for $E_{2\mathcal{NN}}$, $E_{2\mathcal{S}}$, and $E_{2\mathcal{B}}$. We now consider each of the four events in further detail.

4.1. Two nested mutations

In this case the clade subtended by one mutation is a proper subclade of the other [Figure 2(a)]. The genealogy of two nested mutations was also studied by Hobolth and Wiuf (2009), though using a different model of mutation.

Consider the event $E_{2\mathcal{N}}^{(b,c)}$, and assume for now that $a \neq c$, so our observation is of the form $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, with $n_a, n_b, n_c > 0$, and $n_a + n_b + n_c = n$. The sample configuration immediately before the younger mutation event is of the form $\boldsymbol{l}_y = l_y\boldsymbol{e}_a + m\boldsymbol{e}_b + \boldsymbol{e}_c$, and immediately before the older mutation event it is of the form $\boldsymbol{l}_o = l_o\boldsymbol{e}_a + \boldsymbol{e}_b$. For the mutations to be nested we must have $1 \leq m \leq n_b$, and $1 \leq l_o \leq l_y \leq n_a$. Denote a history compatible with these requirements by $\mathcal{H}_{2\mathcal{N}}$; this is a sequence of configurations of the form $\boldsymbol{n} \mapsto \ldots \mapsto \boldsymbol{l}_y \mapsto (\boldsymbol{l}_y - \boldsymbol{e}_c + \boldsymbol{e}_b) \mapsto \ldots \mapsto \boldsymbol{l}_o \mapsto (\boldsymbol{l}_o - \boldsymbol{e}_b + \boldsymbol{e}_a) \mapsto \boldsymbol{e}_a$. Using $\binom{n}{n_a, n_b, n_c}$ to denote the trinomial coefficient, we have the following lemma:

Lemma 4.1. *Conditional on $E_{2\mathcal{N}}^{(b,c)}$, on the observed sample configuration $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, and on the sample configurations $\boldsymbol{l}_y = l_y\boldsymbol{e}_a + m\boldsymbol{e}_b + \boldsymbol{e}_c$ and $\boldsymbol{l}_o = l_o\boldsymbol{e}_a + \boldsymbol{e}_b$ immediately before the times of the two nested mutation events, the distribution of compatible histories $\mathcal{H}_{2\mathcal{N}}$ is* uniform. *It is given by*

$$p(\mathcal{H}_{2\mathcal{N}} \mid E_{2\mathcal{N}}^{(b,c)}, \boldsymbol{n}, \boldsymbol{l}_y, \boldsymbol{l}_o) = \left[\binom{n - l_y - m - 1}{n_a - l_y, n_b - m, n_c - 1} \binom{m + l_y - l_o}{m} \right]^{-1}.$$

Moreover, the (unconditional) probability of such histories is

$$\begin{aligned} p(\mathcal{H}_{2\mathcal{N}}) &= p(\mathcal{H}_{2\mathcal{N}}, E_{2\mathcal{N}}, \boldsymbol{n}, \boldsymbol{l}_y, \boldsymbol{l}_o) \\ &= \theta^2 P_{ab} P_{bc} \frac{(n_a - 1)!(n_b - 1)!(n_c - 1)!}{(1 + \theta)_{n-1}} \\ &\quad \times \frac{m l_o}{(m + l_y + \theta)(l_o + \theta)} \cdot \frac{m + 1}{m + l_y + 1}. \end{aligned} \tag{7}$$

Proof. See Appendix A. □

Summing over compatible histories and over valid combinations of m, l_y,

and l_o, we obtain

$$p(\boldsymbol{n}, E_{2\mathcal{N}}^{(b,c)}) = \sum_{l_y=1}^{n_a} \sum_{l_o=1}^{l_y} \sum_{m=1}^{n_b} \binom{n-l_y-m-1}{n_a-l_y,\ n_b-m,\ n_c-1} \binom{m+l_y-l_o}{m} p(\mathcal{H}_{2\mathcal{N}})$$

$$= \begin{cases} \theta^2 P_{ab} P_{bc} \dfrac{(n-1)!}{(1+\theta)_{n-1}} \displaystyle\sum_{l_y=1}^{n_a} \sum_{l_o=1}^{l_y} \sum_{m=1}^{n_b} \dfrac{\binom{n_a-1}{l_y-1}\binom{n_b-1}{m-1}\binom{m+l_y-l_o}{m}}{\binom{n-1}{m+l_y}\binom{m+l_y}{m+1}} \\ \qquad\qquad \times \dfrac{1}{m+l_y+1} \cdot \dfrac{l_o}{l_o+\theta} \cdot \dfrac{1}{m+l_y+\theta}, \quad (8) \\ \quad \text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c, \text{ where } a, b, c \text{ are all} \\ \quad \text{distinct, and } 1 \le n_a, n_b, n_c \le n; \\ 0, \quad \text{otherwise.} \end{cases}$$

One can relax the age-ordering on mutations by noting that, for $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$,

$$p(\boldsymbol{n}, E_{2\mathcal{N}}) = p(\boldsymbol{n}, E_{2\mathcal{N}}^{(b,c)}) + p(\boldsymbol{n}, E_{2\mathcal{N}}^{(c,b)}).$$

Similar relaxations apply to the remaining cases described below.

Finally, we return to the possibility that $a = c$. We denote this by $E_{2\mathcal{N}}^{(b,a)}$, and have a sample of the form $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b$, with $n_a, n_b > 0$, and $n_a + n_b = n$. Now, n_a comprises gametes whose alleles are truly ancestral and gametes whose alleles are atavistic. We can still however apply the previous result [equation (8)], first by treating a and c as if they were distinct, then summing over all possible values for the number of observed a alleles and the number of observed c alleles such that their sum is held fixed, and finally setting $c = a$ in the resulting expression:

$$p(\boldsymbol{n}, E_{2\mathcal{N}}^{(b,a)}) = \sum_{k=1}^{n_a-1} p(k\boldsymbol{e}_a + n_b\boldsymbol{e}_b + (n-n_b-k)\boldsymbol{e}_c, E_{2\mathcal{N}}^{(b,c)})\big|_{c=a}$$

$$= \begin{cases} \theta^2 P_{ab} P_{ba} \dfrac{(n-1)!}{(1+\theta)_{n-1}} \displaystyle\sum_{l_y=1}^{n_a-1} \sum_{l_o=1}^{l_y} \sum_{m=1}^{n_b} \dfrac{\binom{n_a-1}{l_y}\binom{n_b-1}{m-1}\binom{m+l_y-l_o}{m}}{\binom{n-1}{m+l_y}\binom{m+l_y}{m+1}} \\ \qquad\qquad \times \dfrac{1}{m+l_y+1} \cdot \dfrac{l_o}{l_o+\theta} \cdot \dfrac{1}{m+l_y+\theta}, \\ \quad \text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b, \text{ where } b \ne a, \text{ and } 1 \le n_a, n_b \le n; \\ 0, \quad \text{otherwise.} \end{cases}$$

4.2. Two nonnested mutations

In this case the clades subtended by the two mutations are disjoint [Figure 2(b)]. We also exclude the possibility that the two mutations reside on the basal (innermost) branches of the coalescent tree, which could result in a monomorphic sample. Consider the event $E^{(b,c)}_{2\mathcal{NN}}$, where $b \neq c$, $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, with $n_a, n_b, n_c > 0$, and $n_a + n_b + n_c = n$. Suppose that, immediately before the younger mutation, the sample configuration is $\boldsymbol{l}_y = l_y\boldsymbol{e}_a + m\boldsymbol{e}_b + \boldsymbol{e}_c$, and immediately before the older mutation it is $\boldsymbol{l}_o = l_o\boldsymbol{e}_a + \boldsymbol{e}_b$. Consideration of the genealogy [Figure 2(b)] leads to the following restrictions: $1 \leq m \leq n_b$, $1 \leq l_y \leq n_a$, and $1 \leq l_o \leq l_y + 1$. One can argue as in the previous subsection to obtain the following:

Lemma 4.2. *Conditional on $E^{(b,c)}_{2\mathcal{NN}}$, on the observed sample configuration $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, and on the sample configurations $\boldsymbol{l}_y = l_y\boldsymbol{e}_a + m\boldsymbol{e}_b + \boldsymbol{e}_c$ and $\boldsymbol{l}_o = l_o\boldsymbol{e}_a + \boldsymbol{e}_b$ immediately before the times of the two nonnested mutation events, the distribution of compatible histories $\mathcal{H}_{2\mathcal{NN}}$ is* uniform. *It is given by*

$$p(\mathcal{H}_{2\mathcal{NN}} \mid E^{(b,c)}_{2\mathcal{NN}}, \boldsymbol{n}, \boldsymbol{l}_y, \boldsymbol{l}_o) = \left[\binom{n - m - l_y - 1}{n_a - l_y,\ n_b - m,\ n_c - 1} \binom{m + l_y - l_o}{m - 1} \right]^{-1}.$$

Moreover, the (unconditional) probability of such histories is

$$\begin{aligned} p(\mathcal{H}_{2\mathcal{NN}}) &= p(\mathcal{H}_{2\mathcal{NN}}, E^{(b,c)}_{2\mathcal{NN}}, \boldsymbol{n}, \boldsymbol{l}_y, \boldsymbol{l}_o) \\ &= \theta^2 P_{ab} P_{ac} \frac{(n_a - 1)!(n_b - 1)!(n_c - 1)!}{(1 + \theta)_{n-1}} \\ &\quad \times \frac{l_y l_o}{(l_y + m + \theta)(l_o + \theta)} \cdot \frac{l_y + 1}{l_y + m + 1}. \end{aligned} \tag{9}$$

Using this lemma to sum over compatible histories and over m, l_y, and l_o,

after some simplification we obtain

$$p(\boldsymbol{n}, E^{(b,c)}_{2\mathcal{NN}}) = \sum_{m=1}^{n_b}\sum_{l_y=1}^{n_a}\sum_{l_o=1}^{l_y+1}\binom{n-m-l_y-1}{n_a-l_y,\ n_b-m,\ n_c-1}\binom{m+l_y-l_o}{m-1}p(\mathcal{H}_{2\mathcal{NN}})$$

$$= \begin{cases} \theta^2 P_{ab}P_{ac}\dfrac{(n-1)!}{(1+\theta)_{n-1}}\displaystyle\sum_{m=1}^{n_b}\sum_{l_y=1}^{n_a}\sum_{l_o=1}^{l_y+1}\dfrac{\binom{n_a-1}{l_y-1}\binom{n_b-1}{m-1}\binom{m+l_y-l_o}{m-1}}{\binom{n-1}{m+l_y}\binom{m+l_y}{l_y+1}} \\ \qquad\qquad \times \dfrac{l_o}{l_o+\theta}\cdot\dfrac{1}{m+l_y+1}\cdot\dfrac{1}{m+l_y+\theta}, \\ \quad \text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c, \text{ where } a,b,c \text{ are all} \\ \quad \text{distinct, and } 1 \le n_a, n_b, n_c \le n; \\ 0, \quad \text{otherwise.} \end{cases} \tag{10}$$

Also as before, the result (10) can also be used to handle the special case $b = c$:

$$p(\boldsymbol{n}, E^{(b,b)}_{2\mathcal{NN}}) = \sum_{k=1}^{n_b-1} p(n_a\boldsymbol{e}_a + k\boldsymbol{e}_b + (n-n_a-k)\boldsymbol{e}_c, E^{(b,c)}_{2\mathcal{NN}})\big|_{b=c}$$

$$= \begin{cases} \theta^2 P^2_{ab}\dfrac{(n-1)!}{(1+\theta)_{n-1}}\displaystyle\sum_{m=1}^{n_b-1}\sum_{l_y=1}^{n_a}\sum_{l_o=1}^{l_y+1}\dfrac{\binom{n_a-1}{l_y-1}\binom{n_b-1}{m}\binom{m+l_y-l_o}{m-1}}{\binom{n-1}{m+l_y}\binom{m+l_y}{l_y+1}} \\ \qquad\qquad \times \dfrac{l_o}{l_o+\theta}\cdot\dfrac{1}{m+l_y+1}\cdot\dfrac{1}{m+l_y+\theta}, \\ \quad \text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b, \text{ where } b \ne a, \text{ and } 1 \le n_a, n_b \le n; \\ 0, \quad \text{otherwise.} \end{cases}$$

4.3. Two mutations on the same branch

Here, two mutation events reside on the same edge of the coalescent tree, as illustrated in Figure 2(c). Consider first the subevent $E^{(b,c)}_{2\mathcal{S}}$, with $a \ne c$, so that $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_c\boldsymbol{e}_c$, with $n_a, n_c > 0$, and $n_a + n_c = n$. Suppose the configuration immediately prior to the younger mutation event is $\boldsymbol{l}_y = l_y\boldsymbol{e}_a + \boldsymbol{e}_c$, and immediately prior to the older mutation is $\boldsymbol{l}_o = l_o\boldsymbol{e}_a + \boldsymbol{e}_b$. We argue as before to yield the following:

Lemma 4.3. *Conditional on $E^{(b,c)}_{2\mathcal{S}}$, on the observed sample configuration $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_c\boldsymbol{e}_c$, and on the sample configurations $\boldsymbol{l}_y = l_y\boldsymbol{e}_a + \boldsymbol{e}_c$ and*

$\boldsymbol{l}_o = l_o\boldsymbol{e}_a + \boldsymbol{e}_b$ *immediately before the times of the two mutation events, the distribution of compatible histories* $\mathcal{H}_{2\mathcal{S}}$ *is* uniform. *It is given by*

$$p(\mathcal{H}_{2\mathcal{S}} \mid E_{2\mathcal{S}}^{(b,c)}, \boldsymbol{n}, \boldsymbol{l}_y, \boldsymbol{l}_o) = \binom{n - l_y - 1}{n_c - 1}^{-1}.$$

Moreover, the (unconditional) probability of such histories is

$$\begin{aligned} p(\mathcal{H}_{2\mathcal{S}}) &= p(\mathcal{H}_{2\mathcal{S}}, E_{2\mathcal{S}}^{(b,c)}, \boldsymbol{n}, \boldsymbol{l}_y, \boldsymbol{l}_o) \\ &= \theta^2 P_{ab} P_{bc} \frac{(n_a - 1)!(n_c - 1)!}{(1+\theta)_{n-1}} \cdot \frac{l_o}{l_o + \theta} \cdot \frac{1}{(l_y + 1)(l_y + \theta)}. \end{aligned} \tag{11}$$

Hence, summing over compatible histories and then over l_o and l_y, we obtain

$$\begin{aligned} p(\boldsymbol{n}, E_{2\mathcal{S}}^{(b,c)}) &= \sum_{l_y=1}^{n_a} \sum_{l_o=1}^{l_y} \binom{n - l_y - 1}{n_c - 1} p(\mathcal{H}_{2\mathcal{S}}) \\ &= \begin{cases} \theta^2 P_{ab} P_{bc} \dfrac{(n-1)!}{(1+\theta)_{n-1}} \displaystyle\sum_{l_y=1}^{n_a} \sum_{l_o=1}^{l_y} \frac{\binom{n_a - 1}{l_y - 1}}{\binom{n-1}{l_y}} \cdot \frac{l_o}{l_o + \theta} \cdot \frac{1}{l_y(l_y+1)(l_y+\theta)}, \\ \qquad \text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_c\boldsymbol{e}_c, \text{ where } a, b, c \text{ are all distinct,} \\ \qquad \text{and } 1 \le n_a, n_c \le n; \\ 0, \quad \text{otherwise.} \end{cases} \end{aligned} \tag{12}$$

One can relax the restriction on the unobserved allele being b by summing over each possible b. This is achieved simply by replacing $P_{ab}P_{bc}$ in (12) with $(P^2)_{ac}$.

When $a = c$, the sample must be $\boldsymbol{n} = n\boldsymbol{e}_a$, and repeating earlier arguments we obtain

$$\begin{aligned} p(\boldsymbol{n}, E_{2\mathcal{S}}^{(b,a)}) &= \sum_{k=1}^{n-1} p(k\boldsymbol{e}_a + (n-k)\boldsymbol{e}_c, E_{2\mathcal{S}}^{(b,c)})\big|_{a=c} \\ &= \begin{cases} \theta^2 P_{ab} P_{ba} \dfrac{(n-1)!}{(1+\theta)_{n-1}} \displaystyle\sum_{l_y=1}^{n-1} \sum_{l_o=1}^{l_y} \frac{l_o}{l_o + \theta} \cdot \frac{1}{l_y(l_y+1)(l_y+\theta)}, \\ \qquad \text{if } \boldsymbol{n} = n\boldsymbol{e}_a, \\ 0, \quad \text{otherwise.} \end{cases} \end{aligned}$$

Again, the stipulation on the unobserved allele being b can be relaxed by replacing $P_{ab}P_{ba}$ with $(P^2)_{aa}$.

4.4. Two mutations on the basal branches

Here, the mutations reside on the last two edges in the coalescent tree to exist going back in time, as illustrated in Figure 2(d). Consider first the subevent $E_{2\mathcal{B}}^{(b,c)}$, with $b \neq c$, so that $\boldsymbol{n} = n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, $n_b, n_c > 0$, and $n_b + n_c = n$. Suppose the configuration immediately prior to the younger mutation event is $\boldsymbol{l}_y = m\boldsymbol{e}_b + \boldsymbol{e}_c$. The configuration at the older mutation event must be $\boldsymbol{e}_a + \boldsymbol{e}_b$. Arguing as in previous subsections yields the following:

Lemma 4.4. *Conditional on $E_{2\mathcal{B}}^{(b,c)}$, on the sample configuration $\boldsymbol{n} = n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, and on the sample configuration $\boldsymbol{l}_y = m\boldsymbol{e}_b + \boldsymbol{e}_c$ immediately before the time of the younger mutation event, the distribution of compatible histories $\mathcal{H}_{2\mathcal{B}}$ is* uniform. *It is given by*

$$p(\mathcal{H}_{2\mathcal{B}} \mid E_{2\mathcal{B}}^{(b,c)}, \boldsymbol{l}_y) = \binom{n-m-1}{n_c-1}^{-1}.$$

Moreover, the (unconditional) probability of such histories is

$$\begin{aligned} p(\mathcal{H}_{2\mathcal{B}}) &= p(\mathcal{H}_{2\mathcal{B}}, E_{2\mathcal{B}}^{(b,c)}, \boldsymbol{n}, \boldsymbol{l}_y) \\ &= \frac{\theta^2}{1+\theta} P_{ab}P_{ac} \frac{(n_b-1)!(n_c-1)!}{(1+\theta)_{n-1}} \frac{1}{(m+1)(m+\theta)}. \end{aligned} \tag{13}$$

Hence, summing over compatible histories and over m, we obtain

$$\begin{aligned} p(\boldsymbol{n}, E_{2\mathcal{B}}^{(b,c)}) &= \sum_{m=1}^{n_b} \binom{n-m-1}{n_c-1} p(\mathcal{H}_{2\mathcal{B}}) \\ &= \begin{cases} \dfrac{\theta^2}{1+\theta} P_{ab}P_{ac} \dfrac{(n-1)!}{(1+\theta)_{n-1}} \displaystyle\sum_{m=1}^{n_b} \frac{\binom{n_b-1}{m-1}}{\binom{n-1}{m}} \frac{1}{m(m+1)(m+\theta)}, \\ \qquad \text{if } \boldsymbol{n} = n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c, \text{ where } a, b, c \text{ are all distinct,} \\ \qquad \text{and } 1 \leq n_b, n_c \leq n; \\ 0, \quad \text{otherwise.} \end{cases} \end{aligned} \tag{14}$$

Finally, when $b = c$ the sample must be of the form $\boldsymbol{n} = n_b\boldsymbol{e}_b$, and the probability of observing such a configuration as a result of two mutation

events on the basal branches is given by

$$p(\boldsymbol{n}, E_{2\mathcal{B}}^{(b,b)}) = \sum_{k=1}^{n-1} p(k\boldsymbol{e}_b + (n-k)\boldsymbol{e}_c, E_{2\mathcal{B}}^{(b,c)})\big|_{b=c}$$
$$= \begin{cases} \dfrac{\theta^2}{1+\theta}(P_{ab})^2 \dfrac{(n-1)!}{(1+\theta)_{n-1}} \displaystyle\sum_{m=1}^{n-1} \frac{1}{m(m+1)(m+\theta)}, & \text{if } \boldsymbol{n} = n\boldsymbol{e}_b, \\ 0, & \text{otherwise.} \end{cases}$$

5. The limit $\theta \to 0$

To make further progress, we derive expressions in the limit as $\theta \to 0$. Results are therefore approximate for nonzero θ, but should still exhibit good accuracy when applied to human single-nucleotide polymorphism data for example, for which θ is small, as noted in the introduction.

Our results will be expressed in terms of harmonic numbers, for which we use the following notation:

$$H_n = \sum_{j=1}^{n} \frac{1}{j}, \qquad \text{and} \qquad H_n^{(2)} = \sum_{j=1}^{n} \frac{1}{j^2}.$$

Further, let $c_n^{(s)}$ denote the *sth order generalized harmonic number* (Roman, 1993), defined for $s \geq 0$ and $n \geq 1$ by

$$c_n^{(s)} = \begin{cases} 1, & \text{if } s = 0, \\ \displaystyle\sum_{j=1}^{n} \frac{c_j^{(s-1)}}{j}, & \text{if } s > 0. \end{cases}$$

In particular,

$$c_n^{(1)} = H_n, \quad \text{and} \quad c_n^{(2)} = \sum_{j=1}^{n} \frac{H_j}{j} = \frac{1}{2}\left[(H_n)^2 + H_n^{(2)}\right]. \tag{15}$$

This last identity is easily verified by induction on n. To simplify notation, we also introduce the function

$$d(n_a, n_b, n_c) = \frac{1}{(n_a+n_b)(n_a+n_b-1)}\left[1 + \frac{n}{n_c} - \frac{2n(H_n - H_{n_c-1})}{n_a+n_b+1}\right], \tag{16}$$

where $n = n_a + n_b + n_c$.

Theorem 5.1. *As $\theta \to 0$, the joint probability of observing a particular sample configuration $\boldsymbol{n}$ together with the topological characterization of the genealogy satisfies:*

$$p(\boldsymbol{n}, E_1) = \frac{\theta P_{ab}}{n_b} - \theta^2 P_{ab}\left[\frac{H_{n-1}}{n_b} + \frac{1}{n_a}(H_n - H_{n_b-1})\right] + O(\theta^3), \tag{17}$$
$$\text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b \text{ and } 0 \text{ otherwise,}$$

$$p(\boldsymbol{n}, E_{2\mathcal{N}}^{(b,c)}) = \theta^2 P_{ab}P_{bc}d(n_a, n_b, n_c) + O(\theta^3), \tag{18}$$
$$\text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c \text{ and } 0 \text{ otherwise,}$$

$$p(\boldsymbol{n}, E_{2\mathcal{N}}^{(b,a)}) = \theta^2 P_{ab}P_{ba}\left[\frac{1}{n_b+1}\left[1 - \frac{n}{n_b}(H_{n-1} - H_{n_a-1})\right] + \frac{H_n - 1}{n-1}\right] \tag{19}$$
$$+ O(\theta^3), \qquad \text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b \text{ and } 0 \text{ otherwise,}$$

$$p(\boldsymbol{n}, E_{2\mathcal{N}\mathcal{N}}^{(b,c)}) = \theta^2 P_{ab}P_{ac}\left[\frac{1}{n_c(n_b+n_c)} - d(n_a, n_b, n_c)\right] + O(\theta^3), \tag{20}$$
$$\text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c \text{ and } 0 \text{ otherwise,}$$

$$p(\boldsymbol{n}, E_{2\mathcal{N}\mathcal{N}}^{(b,b)}) = \theta^2 P_{ab}^2\left[\frac{H_{n_b-1}}{n_b} - \frac{1}{n_a+1}\left[1 - \frac{n}{n_a}(H_{n-1} - H_{n_b-1})\right]\right.$$
$$\left. - \frac{H_n - 1}{n-1}\right] + O(\theta^3), \tag{21}$$
$$\text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b \text{ and } 0 \text{ otherwise,}$$

$$p(\boldsymbol{n}, E_{2\mathcal{S}}^{(b,c)}) = \frac{\theta^2 P_{ab}P_{bc}}{n_a}\left[\frac{n}{n_a+1}(H_n - H_{n_c-1}) - 1\right] + O(\theta^3), \tag{22}$$
$$\text{if } \boldsymbol{n} = n_a\boldsymbol{e}_a + n_c\boldsymbol{e}_c \text{ and } 0 \text{ otherwise,}$$

$$p(\boldsymbol{n}, E_{2\mathcal{S}}^{(b,a)}) = \theta^2 P_{ab}P_{ba}\left[1 - \frac{1}{n}\right] + O(\theta^3), \tag{23}$$
$$\text{if } \boldsymbol{n} = n\boldsymbol{e}_a \text{ and } 0 \text{ otherwise,}$$

$$p(\boldsymbol{n}, E_{2\mathcal{B}}^{(b,c)}) = \frac{\theta^2 P_{ab}P_{ac}}{n_b+1}\left[1 - \frac{n_c - 1}{n_b}(H_{n-1} - H_{n_c-1})\right] + O(\theta^3), \tag{24}$$

if $\boldsymbol{n} = n_b \boldsymbol{e}_b + n_c \boldsymbol{e}_c$ *and* 0 *otherwise,*

$$p(\boldsymbol{n}, E_{2\mathcal{B}}^{(b,b)}) = \theta^2 P_{ab}^2 \left[H_{n-1}^{(2)} - 1 + \frac{1}{n} \right] + O(\theta^3), \tag{25}$$

if $\boldsymbol{n} = n\boldsymbol{e}_b$ *and* 0 *otherwise,*

where a, b, c *are all distinct.*

Proof. See Appendix A. □

We illustrate the use of Theorem 5.1 with two simple applications. First, it can be used to calculate the sample frequency spectrum of a site that has undergone two mutations. Figure 3 shows the sample frequency spectrum conditional on E_2, for a diallelic model with alleles 1 and 2, $a = 1$ being ancestral [so $P = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$]. Suppose we have a large quantity of SNP data, which we incorrectly assume to satisfy the infinite sites model. The occasional second mutation will distort the frequency spectrum. Figure 3 shows that if we had used equation (1) to predict the sample frequency spectrum at a site that had in fact undergone two mutation events, then the main effect would be to slightly underestimate the probability that a mutant is seen in moderate to high frequency, and to grossly overestimate the probability that a mutant is at very low frequency. This does not mean, however, that we should expect the sample frequency spectrum of a site undergoing *at most* two mutation events to be strongly affected. An upper bound on this effect can be found by considering the distribution of the number of mutation events (Tavaré, 1984):

$$p(E_s) = \frac{n-1}{\theta} \sum_{j=1}^{n-1} (-1)^{j-1} \binom{n-2}{j-1} \left(\frac{\theta}{j+\theta} \right)^{s+1}. \tag{26}$$

For example, suppose we compare a histogram of allele frequencies from SNP data with the distribution (1). When $\theta = 0.01$, equation (26) tells us that polymorphic sites resulting from more than one mutation will make up at most 2.3% of the area under the histogram. In other words, we expect the effect of recurrent mutation on the sample frequency spectrum of a randomly chosen polymorphic site to be small. For recurrent mutation to have an appreciable effect, it requires a higher mutation rate (see Table 1).

The diallelic model considered above can be seen as a crude approximation of the evolution of a single nucleotide, in which say the transversion

θ	$p(E_0)$	$p(E_1)$	$p(E_2)$	$p(E_{\geq 3})$	$p(E_{\geq 2} \mid E_{\geq 1})$
0.001	0.9958	0.0042	0.0000	0.0000	0.0023
0.01	0.9584	0.0406	0.0009	0.0000	0.0229
0.05	0.8100	0.1691	0.0192	0.0017	0.1099
0.1	0.6586	0.2702	0.0601	0.0111	0.2085

Table 1: The probability that the genealogy for a sample of $n = 40$ gametes contains 0, 1, 2, or more than two mutations, and the probability of more than one mutation conditional on at least one.

rate is zero. A second application of Theorem 5.1 is to investigate more realistic models of sequence evolution, also incorporating uncertainty over the allele of the MRCA. Now let $K = 4$, representing the nucleotides A, G, C, and T. One way to set a general matrix P is to match its entries to empirical estimation of rates of mutation; for illustration we use those reported in Table 1 of Tamura and Nei (1993) for a human mtDNA sequence dataset (normalizing so that each nucleotide has the same overall mutation rate). Suppose our prior probabilities for the identity of the ancestral nucleotide are $(\pi_{\mathrm{A}}, \pi_{\mathrm{G}}, \pi_{\mathrm{C}}, \pi_{\mathrm{T}})$; they may be based on, for example, an outgroup sequence allowing for some probability of error, on the stationary distribution of P, or on empirical base frequencies. Here we use the last option (Tamura and Nei, 1993): $(\pi_{\mathrm{A}}, \pi_{\mathrm{G}}, \pi_{\mathrm{C}}, \pi_{\mathrm{T}}) = (0.321, 0.132, 0.314, 0.233)$. Equation (17) makes explicit how we should use prior information on the distribution of the MRCA. For example, if we observe an A-G polymorphism, then:

$$\begin{aligned} p_u(n_{\mathrm{A}}\boldsymbol{e}_{\mathrm{A}} + n_{\mathrm{G}}\boldsymbol{e}_{\mathrm{G}} \mid E_1, \{n_{\mathrm{A}}, n_{\mathrm{G}} > 0\}) &= \frac{\sum_{a \in \{\mathrm{A},\mathrm{G}\}} \pi_a p(n_a \boldsymbol{e}_a + n_{\bar{a}} \boldsymbol{e}_{\bar{a}}, E_1)}{p(E_1, \{n_{\mathrm{A}}, n_{\mathrm{G}} > 0\})} \\ &= \frac{\pi_{\mathrm{A}} P_{\mathrm{AG}} n_{\mathrm{G}}^{-1} + \pi_{\mathrm{G}} P_{\mathrm{GA}} n_{\mathrm{A}}^{-1}}{H_{n-1}(\pi_{\mathrm{A}} P_{\mathrm{AG}} + \pi_{\mathrm{G}} P_{\mathrm{GA}})} + O(\theta). \end{aligned} \tag{27}$$

The subscript u is used to denote that the allele of the MRCA is unknown here. We sum a over possible ancestral alleles; the mutant allele is denoted $\bar{a}$. The sample frequency spectrum according to (27), ignoring terms of $O(\theta)$, is shown in Figure 4. As is clear from the figure, uncertainty over the MRCA introduces modes at both $n_{\mathrm{G}} = 1$ and $n_{\mathrm{G}} = n-1$, with relative heights determined by $\pi_{\mathrm{A}} P_{\mathrm{AG}}$ and $\pi_{\mathrm{G}} P_{\mathrm{GA}}$. By contrast, a classical approach is equivalent to assuming $\pi_{\mathrm{A}} P_{\mathrm{AG}} = \pi_{\mathrm{G}} P_{\mathrm{GA}}$, leading to a symmetric frequency spectrum;

for this reason it is common to report a folded spectrum by binning values for each $(n_{\mathrm{G}}, n - n_{\mathrm{G}})$ pair, $n_{\mathrm{G}} = 1, \ldots, n-1$. Figure 4 demonstrates how additional prior information on the ancestral allele can be incorporated properly, in this example leading to a shift in favour of the mode at $n_G = 1$ so that the spectrum is no longer symmetric. We also repeated this analysis using equations (19), (21), (22), and (24) to allow for up to *two* mutations to explain a diallelic site, with the sample frequency spectrum left almost unchanged, as discussed above.

We can sum over all possible outcomes for $\boldsymbol{n}$, b, and c in Theorem 5.1 in order to find the probabilities of the four events illustrated in Figure 2.

Theorem 5.2. *The distribution of the number of mutation events, satisfies*

$$p(E_s) = \sum_{j=0}^{\infty} \theta^{s+j} c_{n-1}^{(s+j)} (-1)^j \binom{s+j}{j}. \tag{28}$$

The series converges for $\theta < 1$. Further, as $\theta \to 0$,

$$p(E_{2\mathcal{N}}^{(b,c)}) = P_{ab}P_{bc}p(E_{2\mathcal{N}}) = \theta^2 P_{ab}P_{bc}\left[H_n + \frac{1}{n} - 2\right] + O(\theta^3), \tag{29}$$

$$p(E_{2\mathcal{N}\mathcal{N}}^{(b,c)}) = P_{ab}P_{ac}p(E_{2\mathcal{N}\mathcal{N}}) = \theta^2 P_{ab}P_{ac}\left[\frac{(H_{n-1})^2}{2} - \frac{H_{n-1}^{(2)}}{2} - H_n - \frac{1}{n} + 2\right] + O(\theta^3), \tag{30}$$

$$p(E_{2\mathcal{S}}^{(b,c)}) = P_{ab}P_{bc}p(E_{2\mathcal{S}}) = \theta^2 P_{ab}P_{bc}\left[1 - \frac{1}{n}\right] + O(\theta^3), \tag{31}$$

$$p(E_{2\mathcal{B}}^{(b,c)}) = P_{ab}P_{ac}p(E_{2\mathcal{B}}) = \theta^2 P_{ab}P_{ac}\left[H_{n-1}^{(2)} - 1 + \frac{1}{n}\right] + O(\theta^3). \tag{32}$$

Proof. See Appendix A. □

As an application of these results, one can ask: Conditional on precisely two mutation events, what are the probabilities of the four possible outcomes illustrated in Figure 2? Using equations (28) and (29)–(32) and letting $\theta \to 0$ yields:

$$p(E_{2\mathcal{N}} \mid E_2) = \frac{H_n + \frac{1}{n} - 2}{c_{n-1}^{(2)}}, \quad p(E_{2\mathcal{N}\mathcal{N}} \mid E_2) = 1 - \frac{H_{n-1}^{(2)} + H_n + \frac{1}{n} - 2}{c_{n-1}^{(2)}},$$

$$p(E_{2\mathcal{S}} \mid E_2) = \frac{1 - \frac{1}{n}}{c_{n-1}^{(2)}}, \qquad p(E_{2\mathcal{B}} \mid E_2) = \frac{H_{n-1}^{(2)} - 1 + \frac{1}{n}}{c_{n-1}^{(2)}}.$$

Since $c_{n-1}^{(2)}$ grows like $(\log n)^2$ with increasing n, we have that $p(E_{2\mathcal{N}} \mid E_2)$ declines to zero like $1/\log n$, $p(E_{2\mathcal{S}} \mid E_2)$ and $p(E_{2\mathcal{B}} \mid E_2)$ decline to zero like $1/(\log n)^2$, while $p(E_{2\mathcal{N}\mathcal{N}} \mid E_2)$ slowly approaches 1 (see Figure 5). Using Theorem 5.1, in a similar manner one could find the relative probabilities of these topologies conditional on the data.

6. Observed patterns of polymorphism

We can partition the space of coalescent trees in two ways: either by the topology of the tree, as in Sections 3 and 4, or by the observed pattern of polymorphism. In practice, it is the latter that is important since only these are known. Our next goal is therefore to find expressions for the sample frequency spectrum conditional on *observed* events. Assuming at most two mutations, the only possible observed outcomes are:

- O_1: No variation, with all alleles ancestral. The sample is of the form $\boldsymbol{n} = n\boldsymbol{e}_a$.
- $O_{1\mathcal{S}}$: No variation, but the observed allele differs from that of the MRCA. The sample is of the form $\boldsymbol{n} = n\boldsymbol{e}_b$, where $b \neq a$.
- O_2: A regular diallelic polymorphism, with both the ancestral allele and a mutant allele observed. The sample is of the form $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b$, where $b \neq a$.
- $O_{2\mathcal{S}}$: A diallelic polymorphism in which the ancestral allele is not observed; instead, we see two mutant alleles. The sample is of the form $\boldsymbol{n} = n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, where a, b, c are all distinct.
- O_3: A triallelic polymorphism, with one observed allele ancestral. The sample is of the form $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, where a, b, c are all distinct.

Note that we assume that the allele of the MRCA is known without error. However, if one observed $O_{1\mathcal{S}}$ or $O_{2\mathcal{S}}$ in practice, then another explanation is that the allele of the MRCA inferred from the outgroup is incorrect, and that a substitution has occurred on the lineage between the MRCA of the sample and the outgroup.

Using the superscript T to denote matrix transpose, we have the following theorem:

Theorem 6.1. *As $\theta \to 0$, the joint probability of two mutations occurring and the observed pattern of polymorphism is given by:*

$$p(O_1, E_2) = \theta^2 (P^2)_{aa} \left(1 - \frac{1}{n}\right) + O(\theta^3), \tag{33}$$

$$p(O_{1\mathcal{S}}, E_2) = \theta^2 (PP^T)_{aa} \left(H^{(2)}_{n-1} - 1 + \frac{1}{n}\right) + O(\theta^3), \tag{34}$$

$$\begin{aligned} p(O_2, E_2) = \theta^2 \Bigg[& (P^2)_{aa} \left(H_n + \frac{1}{n} - 2\right) + (1 - (P^2)_{aa}) \left(1 - \frac{1}{n}\right) \\ & + (PP^T)_{aa} \left(\frac{(H_{n-1})^2}{2} - \frac{H^{(2)}_{n-1}}{2} - H_n - \frac{1}{n} + 2\right) \Bigg] + O(\theta^3), \end{aligned} \tag{35}$$

$$p(O_{2\mathcal{S}}, E_2) = \theta^2 (1 - (PP^T)_{aa}) \left(H^{(2)}_{n-1} - 1 + \frac{1}{n}\right) + O(\theta^3), \tag{36}$$

$$\begin{aligned} p(O_3, E_2) = \theta^2 \Bigg[& (1 - (P^2)_{aa}) \left(H_n + \frac{1}{n} - 2\right) \\ & + (1 - (PP^T)_{aa}) \left(\frac{(H_{n-1})^2}{2} - \frac{H^{(2)}_{n-1}}{2} - H_n - \frac{1}{n} + 2\right) \Bigg] + O(\theta^3). \end{aligned} \tag{37}$$

Proof. See Appendix A. □

As above, we may also consider the relative probability of these events *conditional* on precisely two mutations having occurred. Again, this entails normalizing equations (33)–(37) by dividing by $p(E_2) = \theta^2 c^{(2)}_{n-1} + O(\theta^3)$, and then letting $\theta \to 0$. The relative probabilities of these outcomes is illustrated in Figure 6, for a simple 4×4 mutation model with each nondiagonal entry in P equal to $1/3$.

Further variations on the arguments of Theorem 6.1 are possible. For example, one can proceed in a similar vein by summing over the relevant probabilities in Theorem 5.1, in order to find closed-form expressions for the *joint* probability of the above events together with a particular sample configuration, $\boldsymbol{n}$. The resulting expressions are easy to obtain but do not simplify very much, so we do not give them explicitly. There is however one important exception which we now consider in further detail. If we observe a triallelic site then we know that at least two mutations must have taken place, and we would like to know the joint sample frequency spectrum for the number of copies of each of the two mutant alleles.

Theorem 6.2. *As $\theta \to 0$,*

$$p(\boldsymbol{n} \mid O_3) = \frac{1}{C}\Bigg[P_{ab}P_{bc}d(n_a, n_b, n_c) + P_{ac}P_{cb}d(n_a, n_c, n_b) + P_{ab}P_{ac}\left(\frac{1}{n_b n_c} - d(n_a, n_b, n_c) - d(n_a, n_c, n_b)\right)\Bigg], \tag{38}$$

if $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, and 0 otherwise, where $d(n_a, n_b, n_c)$ is given by equation (16), and

$$C = \Bigg[(1 - (P^2)_{aa})\left(H_n + \frac{1}{n} - 2\right) + (1 - (PP^T)_{aa})\left(\frac{(H_{n-1})^2}{2} - \frac{H^{(2)}_{n-1}}{2} - H_n - \frac{1}{n} + 2\right)\Bigg]. \tag{39}$$

Proof. See Appendix A. □

We remark that there is no need to condition on E_2 in Theorem 6.2, since O_3 requires at least two mutations, and more than two mutations occurs with probability $O(\theta^3)$. Furthermore, as we noted in the introduction, Theorem 6.2 is unchanged when we allow $P_{ii} > 0$ for any i, since such a "self"-mutation could not lead to O_3 without additional mutations. This is true of all our results for which we condition on O_3, and so for the remainder of this section and for Section 7 we can drop the constraint that the diagonal of P is zero.

Equation (38) simplifies a great deal when the mutation transition matrix takes on a particular form, which we state in the following corollary.

Corollary 6.1. *Suppose the mutation transition matrix P satisfies (2). Then, as $\theta \to 0$,*

$$p(\boldsymbol{n} \mid O_3) = \frac{P_{ab}P_{ac}}{C n_b n_c}, \tag{40}$$

if $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, and 0 otherwise, where the normalizing constant C is given by (39).

For the remainder of this section we continue to assume (2) holds. Employing similar arguments, we have that, conditional on observing two par-

ticular alleles b and c, the sample frequency spectrum is

$$\lim_{\theta\to 0} p(\boldsymbol{n} \mid O_3, \{n_b, n_c > 0\}) = \frac{(n_b n_c)^{-1}}{(H_{n-1})^2 - H_{n-1}^{(2)}}, \quad \boldsymbol{n} = n_a \boldsymbol{e}_a + n_b \boldsymbol{e}_b + n_c \boldsymbol{e}_c, \tag{41}$$

and the marginal spectrum for a particular mutant allele, given that we have observed it, is

$$\lim_{\theta\to 0} p(n_b \mid O_3, \{n_b > 0\}) = \frac{H_{n-n_b-1}(n_b)^{-1}}{(H_{n-1})^2 - H_{n-1}^{(2)}}, \qquad 1 \le n_b \le n-2. \tag{42}$$

One can check that the normalizing constant is correct by summing (42) over n_b, and using the identity

$$\sum_{j=1}^{n-1} \frac{H_{n-j}}{j} = (H_n)^2 - H_n^{(2)},$$

for $n \ge 2$, which is easily verified by induction on n.

Similarly, the marginal spectrum for a particular allele, given that we have observed it and conditional on the number of copies of the other mutant allele, is

$$\lim_{\theta\to 0} p(n_b \mid O_3, \{n_b > 0\}, n_c) = \frac{(n_b)^{-1}}{H_{n-n_c-1}}, \quad 1 \le n_b \le n - n_c - 1,\ n_c \ge 1. \tag{43}$$

Interestingly, the distribution for n_b is proportional to n_b^{-1} regardless of n_c. More generally, Corollary 6.1 tells us: as $\theta \to 0$, and given that we observe three particular alleles, one of which is the ancestral allele, the sample frequency spectrum for the two mutant alleles is proportional to the inverse of the product of the number of observed copies of each of the two mutant alleles (scaled by the relative rate $P_{ab}P_{ac}$ of appearance of these two alleles). This result is a straightforward generalization of the classical result (1), with some mild conditions on P.

There is another way to arrive at Corollary 6.1 when P satisfies (2), and which immediately generalizes to any number of observed alleles, O_l, $2 \le l \le K$. An l-allele version of (2) is to suppose that $P_{ij} = P_j$ whenever j is an observed mutant allele and i is a different observed mutant allele or the observed ancestral allele. We refer to this condition as *parent-independence among observed alleles*.

Theorem 6.3. *Suppose that we observe l distinct alleles, one of which is the ancestral allele a and the rest are labelled by the set $\Lambda \subseteq \{1, \ldots, K\} \setminus \{a\}$. Further suppose that P is parent-independent among observed alleles. Then, as $\theta \to 0$,*

$$p(\boldsymbol{n} \mid O_l) \propto \prod_{j \in \Lambda} \frac{P_j}{n_j}.$$

Proof. See Appendix A. □

7. The age of a mutant allele

In this section we will be interested in the population limits $n_a/n \to f_a$, $n_b/n \to f_b$, and $n_c/n \to f_c$ as $n \to \infty$, and it is implicit throughout that we let $\theta \to 0$ and that a, b, c are all distinct. We assume that there have been no more than two mutation events in the history of the population, so $f_a + f_b + f_c = 1$. Let $\boldsymbol{f} = (f_a, f_b, f_c)$, and A_b, A_c denote the ages at which mutations occurred that gave rise to alleles b and c respectively. Kimura and Ohta (1973) showed that the expected age of a single mutant allele at frequency f in the population is

$$-\frac{2f}{1-f} \ln f.$$

Griffiths and Tavaré (2003) found the expected age of the younger of two nested mutant alleles to be

$$\mathbb{E}[A_c \mid E_{2\mathcal{N}}^{(b,c)}, \boldsymbol{f}] = -2f_c \frac{(1+f_c)\ln f_c + 2(1-f_c)}{2f_c \ln f_c + (1+f_c)(1-f_c)}. \tag{44}$$

Here, we extend this result to find the expected age of the younger of two *nonnested* mutant alleles. From this it is straightforward to obtain the expected age of the younger allele at a triallelic site, regardless of whether the mutations are nested or nonnested—and indeed regardless of whether we know which of the two mutant alleles is younger.

Theorem 7.1. *When it is known which of two nonnested mutant alleles is*

the younger, the expected age of the younger allele is

$$\mathbb{E}[A_c \mid E^{(b,c)}_{2\mathcal{N}\mathcal{N}}, \boldsymbol{f}] = 2f_c \frac{\left[1 + f_c - \frac{(1-f_c)^2}{f_b}\right] \ln f_c + 2(1-f_c) + \frac{(1-f_c)^3 \ln(f_b+f_c)}{f_b(1-f_b-f_c)}}{\frac{(1-f_c)^3}{f_b+f_c} - 2f_c \ln f_c - (1+f_c)(1-f_c)}. \quad (45)$$

Proof. See Appendix A. □

Let A denote the age of the younger of two mutant alleles at a triallelic site, when we do *not* know which of the two is younger. Our goal now is to compute its expectation given the frequencies of the two mutant alleles in the population. This can be achieved by averaging over the possible topologies that could have given rise to a triallelic site:

$$\begin{aligned}\mathbb{E}[A \mid O_3, \boldsymbol{f}] = {} & \mathbb{E}[A_c \mid E^{(b,c)}_{2\mathcal{N}}, \boldsymbol{f}] p(E^{(b,c)}_{2\mathcal{N}} \mid O_3, \boldsymbol{f}) \\ & + \mathbb{E}[A_c \mid E^{(b,c)}_{2\mathcal{N}\mathcal{N}}, \boldsymbol{f}] p(E^{(b,c)}_{2\mathcal{N}\mathcal{N}} \mid O_3, \boldsymbol{f}) \\ & + \mathbb{E}[A_b \mid E^{(c,b)}_{2\mathcal{N}}, \boldsymbol{f}] p(E^{(c,b)}_{2\mathcal{N}} \mid O_3, \boldsymbol{f}) \\ & + \mathbb{E}[A_b \mid E^{(c,b)}_{2\mathcal{N}\mathcal{N}}, \boldsymbol{f}] p(E^{(c,b)}_{2\mathcal{N}\mathcal{N}} \mid O_3, \boldsymbol{f}). \end{aligned} \quad (46)$$

The expectation in each term on the right-hand side is given by equations (44), (45) and their analogues (interchanging the roles of b and c). The probabilities on the right-hand side are also known, since

$$p(E^{(b,c)}_{2\mathcal{N}} \mid O_3, \boldsymbol{f}) = \frac{p(\boldsymbol{f}, E^{(b,c)}_{2\mathcal{N}})}{p(\boldsymbol{f}, O_3)},$$

and these terms are found by letting $n_b/n \to f_b$ and $n_c/n \to f_c$ while $n \to \infty$ in equations (18), (37), and (38). A similar argument applies for the other three terms. We obtain

$$\begin{aligned} p(E^{(b,c)}_{2\mathcal{N}} \mid O_3, \boldsymbol{f}) &= \frac{P_{ab}P_{bc}}{D(1-f_c)^2}\left(1 + \frac{1}{f_c} + \frac{2\ln f_c}{1-f_c}\right), \\ p(E^{(b,c)}_{2\mathcal{N}\mathcal{N}} \mid O_3, \boldsymbol{f}) &= \frac{P_{ab}P_{ac}}{D}\left[\frac{1}{f_b(f_b+f_c)} - \frac{1}{(1-f_c)^2}\left(1 + \frac{1}{f_c} + \frac{2\ln f_c}{1-f_c}\right)\right], \end{aligned}$$

with similar expressions for $p(E_{2\mathcal{N}}^{(c,b)} \mid O_3, \boldsymbol{f})$ and $p(E_{2\mathcal{N}\mathcal{N}}^{(c,b)} \mid O_3, \boldsymbol{f})$, where

$$\begin{aligned} D &= p(E_{2\mathcal{N}}^{(b,c)} \mid O_3, \boldsymbol{f}) + p(E_{2\mathcal{N}}^{(c,b)} \mid O_3, \boldsymbol{f}) + p(E_{2\mathcal{N}\mathcal{N}}^{(b,c)} \mid O_3, \boldsymbol{f}) + p(E_{2\mathcal{N}\mathcal{N}}^{(c,b)} \mid O_3, \boldsymbol{f}) \\ &= \frac{P_{ab}(P_{bc} - P_{ac})}{(1-f_c)^2}\left(1 + \frac{1}{f_c} + \frac{2\ln f_c}{1-f_c}\right) + \frac{P_{ac}(P_{cb} - P_{ab})}{(1-f_b)^2}\left(1 + \frac{1}{f_b} + \frac{2\ln f_b}{1-f_b}\right) \\ &\quad + \frac{P_{ab}P_{ac}}{f_b f_c}. \end{aligned}$$

Substituting these expressions, along with (44) and (45), into (46), we obtain the following:

Theorem 7.2. *The expected age of the younger of two mutant alleles at a triallelic site is*

$$\begin{aligned} \mathbb{E}[A \mid O_3, \boldsymbol{f}] = {} & \frac{2P_{ab}P_{ac}}{D(1-f_b-f_c)}\left(\frac{1}{f_b} + \frac{1}{f_c}\right)\ln(f_b+f_c) \\ & + \frac{2P_{ab}}{D}\left[(P_{ac} - P_{bc})\frac{1+f_c}{(1-f_c)^3} - \frac{P_{ac}}{f_b(1-f_c)}\right]\ln f_c \\ & + \frac{2P_{ac}}{D}\left[(P_{ab} - P_{cb})\frac{1+f_b}{(1-f_b)^3} - \frac{P_{ab}}{f_c(1-f_b)}\right]\ln f_b \\ & + \frac{4P_{ab}(P_{ac} - P_{bc})}{D(1-f_c)^2} + \frac{4P_{ac}(P_{ab} - P_{cb})}{D(1-f_b)^2}. \end{aligned}$$

When P additionally satisfies (2) for the observed alleles a, b, and c, this expression simplifies to

$$\mathbb{E}[A \mid O_3, \boldsymbol{f}] = -\frac{2f_b}{1-f_b}\ln f_b - \frac{2f_c}{1-f_c}\ln f_c + \frac{2(f_b+f_c)}{1-(f_b+f_c)}\ln(f_b+f_c). \quad (47)$$

This curious result, for mutation models satisfying (2), tells us that the mean age of the younger of two mutant alleles at a triallelic site is equal to the sum of the mean ages of two independent mutations at frequencies f_b and f_c, minus the mean age of a single mutant at frequency $f_b + f_c$. Equation (47) is plotted in Figure 8, for various values of f_b and f_c.

Unfortunately, a corresponding expression for the expected age of the older mutation is not analytically tractable, even if we restrict our attention to nested mutations. Hobolth and Wiuf (2009) outline a method of numerical approximation which could also be adapted for nonnested mutations, but we do not pursue this here.

8. Accuracy

It would be interesting to investigate the accuracy of the expressions given in the previous section. For simplicity, we focus on equation (41), and we assume a simple Jukes-Cantor model of mutation in which $K = 4$, and the daughter allele of each mutation is equally likely, so that off-diagonal entries of P are all $1/3$. We wrote a program to solve numerically the system of equations defined by (3), in order to obtain exact results for the sample frequency spectrum for nonzero θ. By solving this system for all sample configurations of a given size, we could calculate exact numerical values of $p(\boldsymbol{n} \mid \{n_b, n_c > 0\})$ for each $\boldsymbol{n}$. We measured the accuracy of equation (41) by its *unsigned relative error*:

$$\left| \frac{\left[\lim_{\theta \to 0} p(\boldsymbol{n} \mid O_3, \{n_b, n_c > 0\})\right] - p(\boldsymbol{n} \mid \{n_b, n_c > 0\})}{p(\boldsymbol{n} \mid \{n_b, n_c > 0\})} \right| \times 100\%,$$

for each $\boldsymbol{n}$. Errors in (41) are a consequence of that fact that in reality θ is nonzero and that there may have been more than two mutation events giving rise to the triallelic sample.

For a given sample size and mutation rate, we summarize the discrepancy between the estimated and actual sample frequency spectrum by the largest relative error across all configurations. Results are summarized in Table 2.

As is clear from the table, accuracy diminishes with increasing θ and also diminishes modestly with increasing n. For application to human SNPs, in which generally $0.001 \leq \theta \leq 0.01$, equation (41) provides an excellent approximation to the sample frequency spectrum. For comparison, Table 2 also shows the maximum relative error incurred when we use the analogous result from Theorem 6.3 for a quadrallelic polymorphism under parent-independence among observed alleles:

$$\lim_{\theta \to 0} p(\boldsymbol{n} \mid O_4, \{n_b, n_c, n_d > 0\}) \propto \frac{1}{n_b n_c n_d}, \tag{48}$$

where $\boldsymbol{n} = n_a \boldsymbol{e}_a + n_b \boldsymbol{e}_b + n_c \boldsymbol{e}_c + n_d \boldsymbol{e}_d$, and a, b, c, d, are all distinct. It should be noted that the relative error is dependent on the sample configuration. Figure 7 shows the relative error incurred by using (48) for a representative slice of quadrallelic sample configurations, and, as is evident, the size of the unsigned relative error approaches its maximum across configurations near

Triallelic						
	θ					
n	0.001	0.01	0.05	0.1	0.5	1.0
10	0.07	0.72	3.52	6.89	30.21	57.02
20	0.13	1.25	6.15	12.04	53.74	112.16
30	0.17	1.61	7.66	14.89	64.78	143.49
40	0.25	2.45	10.57	17.97	70.05	162.82
50	0.35	3.30	13.81	22.81	72.20	175.36
60	0.44	4.16	16.87	27.17	72.52	187.79

Quadrallelic						
	θ					
n	0.001	0.01	0.05	0.1	0.5	1.0
10	0.11	1.10	5.24	9.92	35.93	82.38
20	0.25	2.45	11.16	20.05	67.95	181.83
30	0.39	3.75	16.30	28.01	90.50	266.17
40	0.52	5.01	20.86	34.51	108.49	342.24
50	0.66	6.22	24.92	39.90	123.70	412.73
60	0.79	7.40	28.58	44.47	136.98	479.09

Table 2: Maximum unsigned relative error (%) across configurations, of equation (41) (*top*) and equation (48) (*bottom*), for various samples sizes, n, and mutation rates, θ.

the boundary $n_a = 1$. When $n = 20$ and $\theta = 0.01$ the maximum unsigned relative error of 2.45% is attained at $(n_a, n_b, n_c, n_d) = (1, 6, 6, 7)$. For samples containing more than one copy of the ancestral allele, the size of the relative error can be substantially less than its maximum, as Figure 7 confirms. We obtained qualitatively similar patterns for other slices, not shown.

Errors in the sample frequency spectrum will also cause errors in its application, such as in the estimation of mutation rates and in tests of neutrality. We explore this issue by examining the use of the frequency spectrum to estimate θ. Here we assume a diallelic model ($K = 2$) in which each mutation toggles an allele between its ancestral and mutant state, as we did for Figure 3. Suppose we have L sites, and record the counts of the number of sites with zero, one, ..., n mutant alleles. There are several ways to combine these counts to define an estimator of θ (Achaz, 2009); we focus on S_n, the number of sites observed to be segregating, and ξ_1, the number of sites observed to be

singleton mutants [having configuration $\boldsymbol{n} = (n-1, 1)$]. Assuming at most one mutation event per site, moment estimates using these statistics follow from Watterson (1975) and Fu (1995):

$$\mathbb{E}(S_n) = L\theta H_{n-1}, \tag{49}$$

$$\mathbb{E}(\xi_1) = L\theta, \tag{50}$$

where θ is the population-scaled mutation rate per site. These moments can be corrected to allow for up to two mutation events per site. We have

$$\begin{aligned}\mathbb{E}(S_n) &= L\sum_{i=1}^{n-1} p((n-i, i)), &(51)\\ &= L[1 - p(E_0) - p(E_{2\mathcal{S}}) - p(E_{2\mathcal{B}})] + O(\theta^3),\\ &= L\left[\theta H_{n-1} - \frac{\theta^2}{2}[(H_{n-1})^2 + 3H_{n-1}^{(2)}]\right] + O(\theta^3), &(52)\end{aligned}$$

using (28), (31) and (32). If in (52) we drop terms of $O(\theta^3)$, then we obtain a quadratic equation which can be solved to yield θ in terms of $\mathbb{E}(S_n)/L$. Replacing this expectation with the observed quantity provides a point estimate of θ. A comparison of this method of estimation with the classical estimate from (49) is given in Figure 9. For comparison, we found the *true* value of $\mathbb{E}(S_n)/L$ as a function of θ, allowing any number of mutation events, by using our program again to solve the system (3) numerically over a grid of θ-values and plugging the results into (51). This numerical estimate of the relationship between θ and $\mathbb{E}(S_n)/L$ is also plotted on Figure 9. Comparison of this curve with (49) and (52) shows that when few sites are segregating, say $s_n/L < 0.1$, it is reasonable to assume at most one mutation event per site, whereas for a higher fraction of segregating sites, say $s_n/L < 0.25$, assuming at most two mutation events per site is still reasonable. When a substantial fraction of sites are segregating, neither assumption is accurate. SNP densities in humans typically have $s_n/L < 0.05$, so errors from using (49) or (52) will not usually be serious. However, this fraction will only grow in the near future with increasing sample sizes, and regions of the genome of high diversity in cutting-edge datasets already exceed this level (The 1000 Genomes Project Consortium, 2010, Supplementary Figure 3). A similar calculation was performed for $\mathbb{E}(\xi_1)$:

$$\mathbb{E}(\xi_1) = Lp((n-1, 1)) = L\left[\theta - \theta^2\left(H_{n-1} + \frac{3}{2(n-1)}\right)\right] + O(\theta^3), \tag{53}$$

with qualitatively very similar results (Figure 10).

9. Discussion

We have studied the effect of a second mutation on the sample frequency spectrum of a segregating site, under a model of mutation in which the mutation rate is independent of the current allele but the transitions between alleles are otherwise arbitrary. The problem is made tractable by conditioning on whether or not the two mutations are nested in the genealogy, and as a bonus we also obtain the relative probabilities of these topological events. Other key results include the joint sample frequency spectrum of the two mutant alleles at a triallelic site, and the mean age of the younger of the two alleles in the population. These results take on a particularly simple form when we impose mild additional conditions on P, namely (2): Then, the sample frequency spectrum (1) generalizes to $\propto (n_b n_c)^{-1}$, and the expected age of the younger mutant is a linear combination of the result for single mutants [equation (47)]. It would be interesting to obtain a more intuitive argument for this formula.

At present, several large-scale projects for various species are under way to resequence the genomes of many individuals (hundreds to thousands) in a population. Hence, it may soon become possible to include triallelic polymorphisms in population genomic studies. Indeed, triallelic sites are becoming interesting objects of study in their own right (Hodgkinson and Eyre-Walker, 2010). We believe that the theoretical results presented in this paper should prove useful in that regard.

Acknowledgements

This research is supported in part by an NIH grant R01-GM094402, an Alfred P. Sloan Research Fellowship, and a Packard Fellowship for Science and Engineering.

Appendix A. Proofs of main results

Proof of Lemma 3.1. We argue by writing down the probability of a compatible history using (4), and observe that it depends only on $\boldsymbol{n}$ and $\boldsymbol{l}$. It is clear that for any polymorphic sample whose history is explained by precisely one mutation event, *only* configurations of the form $l_a \boldsymbol{e}_a + \boldsymbol{e}_b$ are possible at the

time of this event. So as we trace the history back in time, we must observe $n_b - 1$ coalescent events of type b alleles and $n_a - l_a$ coalescent events of type a alleles (in some interspersed order), followed by a mutation event taking $l_a \boldsymbol{e}_a + \boldsymbol{e}_b \mapsto (l_a + 1)\boldsymbol{e}_a$, followed by l_a coalescent events of the remaining type a alleles. Think of the history as unwrapping a particular path back through the recursion (4). By multiplying together the coefficients accumulated at each transition, we obtain the probability of this history. Regardless of the order of the interspersed events, this product is

$$\frac{(n_a - 1)(n_a - 2)\dots(l_a)(n_b - 1)!}{(n - 1 + \theta)(n - 2 + \theta)\dots(l_a + 1 + \theta)} \times \frac{\theta P_{ab}}{l_a + \theta} \frac{l_a + 1}{l_a + 1} \times \frac{l_a!}{(l_a + \theta)(l_a - 1 + \theta)\dots(1 + \theta)}.$$

Simplifying, we get (5), which indeed depends only on n_a, n_b, and l_a. There are $\binom{n_a - l_a + n_b - 1}{n_b - 1}$ ways to arrange the first $n_a - l_a + n_b - 1$ events, and thus $\binom{n_a - l_a + n_b - 1}{n_b - 1} = \binom{n - l_a - 1}{n_b - 1}$ such histories. □

Proof of Lemma 4.1. We argue in a similar fashion to Lemma 3.1. Any compatible history must exhibit the following order of events:

- an interspersed collection of $n_a - l_y$ coalescence events of type a alleles, $n_b - m$ coalescence events of type b alleles, and $n_c - 1$ coalescence events of type c alleles,
- a mutation event taking $\boldsymbol{l}_y \mapsto l_y \boldsymbol{e}_a + (m + 1)\boldsymbol{e}_b$,
- an interspersed collection of $l_y - l_o$ coalescence events of type a alleles and m coalescence events of type b alleles,
- a mutation event taking $\boldsymbol{l}_o \mapsto (l_o + 1)\boldsymbol{e}_a$, followed by
- l_o coalescence events of type a alleles.

Regardless of the relative ordering of events within each of these collections,

the product of transition probabilities from (4) is

$$\frac{(n_a-1)(n_a-2)\dots(l_y)(n_b-1)(n_b-2)\dots(m)(n_c-1)!}{(n-1+\theta)(n-2+\theta)\dots(m+l_y+1+\theta)}$$
$$\times\frac{\theta P_{bc}}{m+l_y+\theta}\frac{m+1}{m+l_y+1}\times\frac{(l_y-1)(l_y-2)\dots(l_o)m!}{(m+l_y+\theta)(m+l_y-1+\theta)\dots(l_o+1+\theta)}$$
$$\times\frac{\theta P_{ab}}{l_o+\theta}\frac{l_o+1}{l_o+1}\times\frac{l_o!}{(l_o+\theta)(l_o-1+\theta)\dots(1+\theta)}.$$

Simplifying, we get (7), which is independent of the history except through $\boldsymbol{n}$, $\boldsymbol{l}_y$, and $\boldsymbol{l}_o$. There are $\binom{n-l_y-m-1}{n_a-l_y,n_b-m,n_c-1}$ ways to arrange the first collection of coalescence events, and $\binom{m+l_y-l_o}{m}$ ways to arrange the second collection of coalescence events, so there are $\binom{n-l_y-m-1}{n_a-l_y,n_b-m,n_c-1}\binom{m+l_y-l_o}{m}$ such histories. □

Proof of Lemmas 4.2, 4.3, and 4.4. These are very similar to the proof of Lemma 4.1 and so are omitted. □

Proof of Theorem 5.1. The proof for each of the expressions is the same; we expand the denominator and collect the dominant terms in θ. We make use of the following identity:

$$(1+\theta)_{n-1}=\frac{(\theta)_n}{\theta}=\sum_{k=1}^{n}s(n,k)\theta^{k-1},$$

where $s(n,k)$ are the unsigned Stirling numbers of the first kind. Note also that

$$\begin{aligned}s(n,1)&=(n-1)!,\\ s(n,2)&=(n-1)!H_{n-1},\\ s(n,3)&=\frac{1}{2}(n-1)![(H_{n-1})^2-H_{n-1}^{(2)}].\end{aligned}$$

We will also make use of standard identities for summing over binomial coefficients, and one nonstandard one:

$$\sum_{k=1}^{n}\frac{1}{k}\binom{n-k}{i-1}=\binom{n}{i-1}(H_n-H_{i-1}),\tag{A.1}$$

for $1\le i\le n$. Equation (A.1) is proven by induction by Fu (1995, equation (33)), and using another method by Griffiths (2003, Appendix B).

Expanding (6):

$$
\begin{aligned}
p(\boldsymbol{n}, E_1) &= \theta P_{ab}\left[1-\theta\frac{s(n,2)}{s(n,1)}+O(\theta^2)\right]\sum_{l_a=1}^{n_a}\frac{\binom{n_a-1}{l_a-1}}{\binom{n-1}{l_a}}\cdot\frac{1}{l_a}\left[1-\frac{\theta}{l_a}+O(\theta^2)\right],\\
&= \frac{\theta-\theta^2 H_{n-1}}{n_a}P_{ab}\sum_{l_a=1}^{n_a}\frac{\binom{n_a}{l_a}}{\binom{n-1}{l_a}}-\frac{\theta^2}{n_a}P_{ab}\sum_{l_a=1}^{n_a}\frac{\binom{n_a}{l_a}}{\binom{n-1}{l_a}}\cdot\frac{1}{l_a}+O(\theta^3),\\
&= \frac{\theta-\theta^2 H_{n-1}}{n_a}P_{ab}\sum_{l_a=1}^{n_a}\frac{\binom{n-1-l_a}{n_b-1}}{\binom{n-1}{n_a}}-\frac{\theta^2}{n_a}P_{ab}\sum_{l_a=1}^{n_a}\frac{\binom{n-1-l_a}{n_b-1}}{\binom{n-1}{n_a}}\cdot\frac{1}{l_a}+O(\theta^3),\\
&= \frac{\theta-\theta^2 H_{n-1}}{n_b}P_{ab}-\frac{\theta^2}{n_a}P_{ab}\frac{\binom{n-1}{n_b-1}}{\binom{n-1}{n_a}}(H_n-H_{n_b-1})+O(\theta^3), \qquad \text{by (A.1)},
\end{aligned}
$$

which simplifies to (17), as required. Next, expanding (8):

$$
\begin{aligned}
&p(\boldsymbol{n}, E^{(b,c)}_{2\mathcal{N}}) =\\
&\quad\theta^2 P_{ab}P_{bc}\sum_{l_y=1}^{n_a}\sum_{l_o=1}^{l_y}\sum_{m=1}^{n_b}\frac{\binom{n_a-1}{l_y-1}\binom{n_b-1}{m-1}\binom{m+l_y-l_o}{m}}{\binom{n-1}{m+l_y}\binom{m+l_y}{m+1}}\frac{1}{(m+l_y)(m+l_y+1)}+O(\theta^3),\\
&=\theta^2 P_{ab}P_{bc}\sum_{l_y=1}^{n_a}\sum_{m=1}^{n_b}\frac{\binom{n_a-1}{l_y-1}\binom{n_b-1}{m-1}}{\binom{n-1}{m+l_y}}\cdot\frac{1}{(m+l_y)(m+l_y+1)}+O(\theta^3),\\
&=\frac{\theta^2 P_{ab}P_{bc}}{n-1}\sum_{k=3}^{n_a+n_b+1}\sum_{l_y=1}^{k-2}\frac{1}{k}\frac{\binom{n_a-1}{l_y-1}\binom{n_b-1}{k-l_y-2}}{\binom{n-2}{k-2}}+O(\theta^3), \qquad (k=m+l_y+1),\\
&=\frac{\theta^2 P_{ab}P_{bc}}{n-1}\sum_{k=3}^{n_a+n_b+1}\frac{1}{k}\binom{n_a+n_b-2}{k-3}\binom{n-2}{k-2}^{-1}+O(\theta^3), \qquad (A.2)\\
&=\frac{\theta^2 P_{ab}P_{bc}}{n-1}\sum_{k=3}^{n_a+n_b+1}\frac{k-2}{k}\binom{n-k}{n_c-1}\binom{n-2}{n_c}^{-1}\frac{1}{n_c}+O(\theta^3),\\
&=\frac{\theta^2 P_{ab}P_{bc}}{n-1}\binom{n-2}{n_c}^{-1}\frac{1}{n_c}\left[\binom{n-2}{n_c}-2\sum_{k=1}^{n}\frac{1}{k}\binom{n-k}{n_c-1}\right.\\
&\qquad\qquad\left.+2\binom{n-1}{n_c-1}+\binom{n-2}{n_c-1}\right]+O(\theta^3),\\
&=\frac{\theta^2 P_{ab}P_{bc}}{n-1}\left[\frac{1}{n_c}-2\binom{n}{n_c-1}\binom{n-2}{n_c}^{-1}\frac{1}{n_c}(H_n-H_{n_c-1})\right.
\end{aligned}
$$

$$+\frac{2(n-1)}{(n_a+n_b)(n_a+n_b-1)}+\frac{1}{n_a+n_b-1}\Bigg]+O(\theta^3), \tag{A.3}$$

where the last equality uses (A.1). On rearranging, we recover (18) as required. Note that equation (A.2) is consistent with a result of Hobolth and Wiuf (2009, equation (23). Their expression (24) seems to contain an error; the summation should be over $3,\ldots,n$ rather than $3,\ldots,n-n_b+1$.) Next, expanding (10):

$$\begin{aligned}
&p(\boldsymbol{n}, E^{(b,c)}_{2\mathcal{N}\mathcal{N}}) = \\
&\quad\theta^2 P_{ab}P_{ac}\sum_{m=1}^{n_b}\sum_{l_y=1}^{n_a}\sum_{l_o=1}^{l_y+1}\frac{\binom{n_a-1}{l_y-1}\binom{n_b-1}{m-1}\binom{m+l_y-l_o}{m-1}}{\binom{n-1}{m+l_y}\binom{m+l_y}{l_y+1}}\cdot\frac{1}{(m+l_y)(m+l_y+1)}+O(\theta^3),\\
&=\theta^2 P_{ab}P_{ac}\sum_{m=1}^{n_b}\sum_{l_y=1}^{n_a}\frac{\binom{n_a-1}{l_y-1}\binom{n_b-1}{m-1}}{\binom{n-1}{m+l_y}}\cdot\frac{l_y+1}{m}\cdot\frac{1}{(m+l_y)(m+l_y+1)}+O(\theta^3),\\
&=\frac{\theta^2 P_{ab}P_{ac}}{n-1}\sum_{k=3}^{n_a+n_b+1}\sum_{m=1}^{k-2}\frac{\binom{n_a-1}{k-m-2}\binom{n_b-1}{m-1}}{\binom{n-2}{k-2}}\left(\frac{1}{m}-\frac{1}{k}\right)+O(\theta^3), \qquad (A.4)\\
&=\frac{\theta^2 P_{ab}P_{ac}}{n-1}\sum_{k=3}^{n_a+n_b+1}\sum_{m=1}^{k-2}\frac{\binom{n_a-1}{k-2-m}}{\binom{n-2}{k-2}}\left[\binom{n_b}{m}\frac{1}{n_b}-\binom{n_b-1}{m-1}\frac{1}{k}\right]+O(\theta^3),\\
&=\frac{\theta^2 P_{ab}P_{ac}}{n-1}\sum_{k=3}^{n_a+n_b+1}\Bigg[\binom{n_a+n_b-1}{k-2}\frac{1}{n_b}-\binom{n_a-1}{k-2}\frac{1}{n_b}\\
&\qquad\qquad-\binom{n_a+n_b-2}{k-3}\frac{1}{k}\Bigg]\binom{n-2}{k-2}^{-1}+O(\theta^3), \qquad (A.5)\\
&=\frac{\theta^2 P_{ab}P_{ac}}{n-1}\sum_{k=3}^{n_a+n_b+1}\left[\frac{\binom{n-k}{n_c-1}}{\binom{n-2}{n_a+n_b-1}}\frac{1}{n_b}-\frac{\binom{n-k}{n_b+n_c-1}}{\binom{n-2}{n_a-1}}\frac{1}{n_b}-\binom{n_a+n_b-2}{k-3}\frac{1}{k}\right]+O(\theta^3),\\
&=\frac{\theta^2 P_{ab}P_{ac}}{n-1}\left[\frac{n-1}{n_c(n_b+n_c)}-\sum_{k=3}^{n_a+n_b+1}\binom{n_a+n_b-2}{k-3}\binom{n-2}{k-2}^{-1}\frac{1}{k}\right]+O(\theta^3).
\end{aligned}$$

Finally, apply the same equality relating equations (A.2) and (A.3) to recover (20) as required. Next, expanding (12) and summing over l_o:

$$p(\boldsymbol{n}, E^{(b,c)}_{2\mathcal{S}}) = \theta^2 P_{ab}P_{bc}\sum_{l_y=1}^{n_a}\binom{n_a-1}{l_y-1}\binom{n-1}{l_y}^{-1}\frac{1}{l_y(l_y+1)}+O(\theta^3), \tag{A.6}$$

$$= \frac{\theta^2 P_{ab} P_{bc}}{n_a} \binom{n-1}{n_a}^{-1} \sum_{l_y=2}^{n_a+1} \frac{1}{l_y} \binom{n-l_y}{n_c-1} + O(\theta^3),$$

$$= \frac{\theta^2 P_{ab} P_{bc}}{n_a} \binom{n-1}{n_a}^{-1} \left[\binom{n}{n_c-1} (H_n - H_{n_c-1}) - 1 \right] + O(\theta^3),$$

where the last equality uses (A.1). This simplifies to (22). Finally, expanding (14):

$$p(\boldsymbol{n}, E_{2\mathcal{B}}^{(b,c)}) =$$

$$\theta^2 P_{ab} P_{ac} \sum_{m=1}^{n_b} \binom{n_b-1}{m-1} \binom{n-1}{m}^{-1} \frac{1}{m^2(m+1)} + O(\theta^3), \tag{A.7}$$

$$= \frac{\theta^2 P_{ab} P_{ac}}{n_c} \binom{n-1}{n_b-1}^{-1} \sum_{m=1}^{n_b} \binom{n-1-m}{n_c-1} \left(\frac{1}{m} - \frac{1}{m+1} \right) + O(\theta^3),$$

$$= \frac{\theta^2 P_{ab} P_{ac}}{n_c} \binom{n-1}{n_b-1}^{-1} \left[\sum_{m=1}^{n_b} \binom{n-1-m}{n_c-1} \frac{1}{m} - \sum_{m=2}^{n_b+1} \binom{n-m}{n_c-1} \frac{1}{m} \right] + O(\theta^3),$$

$$= \frac{\theta^2 P_{ab} P_{ac}}{n_c} \binom{n-1}{n_b-1}^{-1} \left[\binom{n-1}{n_c-1} (H_{n-1} - H_{n_c-1}) \right.$$

$$\left. - \binom{n}{n_c-1} (H_n - H_{n_c-1}) + \binom{n-1}{n_c-1} \right] + O(\theta^3),$$

applying (A.1) to each sum in the penultimate expression. This then simplifies to (24). Expressions for $p(\boldsymbol{n}, E_{2\mathcal{N}}^{(b,a)})$, $p(\boldsymbol{n}, E_{2\mathcal{N}\mathcal{N}}^{(b,b)})$, $p(\boldsymbol{n}, E_{2\mathcal{S}}^{(b,a)})$, and $p(\boldsymbol{n}, E_{2\mathcal{B}}^{(b,b)})$ are obtained in a very similar manner and we omit the details. □

Proof of Theorem 5.2. By expanding the denominator of (26) for $\theta < 1$ and applying the following identity (Roman, 1993):

$$c_n^{(s)} = \sum_{j=1}^{n} \binom{n}{j} \frac{(-1)^{j-1}}{j^s},$$

we obtain (28). For the remaining results, we sum over all possible observations $\boldsymbol{n}$ consistent with the event of interest:

$$p(E_{2\mathcal{N}}^{(b,c)}) = \sum_{n_c=1}^{n-2} \sum_{n_b=1}^{n-n_c-1} p((n-n_b-n_c)\boldsymbol{e}_a + n_b \boldsymbol{e}_b + n_c \boldsymbol{e}_c, E_{2\mathcal{N}}^{(b,c)}),$$

$$
\begin{aligned}
&= \frac{\theta^2 P_{ab}P_{bc}}{n-1}\sum_{n_c=1}^{n-2}\sum_{n_b=1}^{n-n_c-1}\sum_{k=3}^{n-n_c+1}\frac{1}{k}\binom{n-n_c-2}{k-3}\binom{n-2}{k-2}^{-1}+O(\theta^3),\\
&= \frac{\theta^2 P_{ab}P_{bc}}{n-1}\sum_{n_c=1}^{n-2}(n-n_c-1)\sum_{k=3}^{n-n_c+1}\frac{1}{k}\binom{n-n_c-2}{k-3}\binom{n-2}{k-2}^{-1}+O(\theta^3),\\
&= \frac{\theta^2 P_{ab}P_{bc}}{n-1}\sum_{k=3}^{n}\sum_{n_c=1}^{n+1-k}\frac{k-2}{k}\binom{n-n_c-1}{k-2}\binom{n-2}{k-2}^{-1}+O(\theta^3),\\
&= \theta^2 P_{ab}P_{bc}\sum_{k=3}^{n}\frac{k-2}{k(k-1)}+O(\theta^3),
\end{aligned}
$$

which simplifies to (29). The second equality above uses (A.2). Continue in this way for the remaining events. Using (A.4):

$$
\begin{aligned}
p(E_{2\mathcal{N}\mathcal{N}}^{(b,c)}) &= \sum_{n_c=1}^{n-2}\sum_{n_b=1}^{n-n_c-1} p((n-n_b-n_c)\boldsymbol{e}_a+n_b\boldsymbol{e}_b+n_c\boldsymbol{e}_c, E_{2\mathcal{N}\mathcal{N}}^{(b,c)}),\\
&= \frac{\theta^2 P_{ab}P_{ac}}{n-1}\sum_{n_c=1}^{n-2}\sum_{n_b=1}^{n-n_c-1}\sum_{k=3}^{n-n_c+1}\sum_{m=1}^{k-2}\frac{\binom{n-n_b-n_c-1}{k-m-2}\binom{n_b-1}{m-1}}{\binom{n-2}{k-2}}\left(\frac{1}{m}-\frac{1}{k}\right)+O(\theta^3),\\
&= \frac{\theta^2 P_{ab}P_{ac}}{n-1}\sum_{n_c=1}^{n-2}\sum_{k=3}^{n-n_c+1}\sum_{m=1}^{k-2}\frac{\binom{n-n_c-1}{k-2}}{\binom{n-2}{k-2}}\left(\frac{1}{m}-\frac{1}{k}\right)+O(\theta^3),\\
&= \frac{\theta^2 P_{ab}P_{ac}}{n-1}\sum_{k=3}^{n}\sum_{m=1}^{k-2}\sum_{n_c=1}^{n+1-k}\frac{\binom{n-n_c-1}{k-2}}{\binom{n-2}{k-2}}\left(\frac{1}{m}-\frac{1}{k}\right)+O(\theta^3),\\
&= \theta^2 P_{ab}P_{ac}\sum_{k=3}^{n}\sum_{m=1}^{k-2}\left(\frac{1}{m}-\frac{1}{k}\right)\frac{1}{k-1}+O(\theta^3),\\
&= \theta^2 P_{ab}P_{ac}\sum_{k=3}^{n}\left(H_{k-2}-\frac{k-2}{k}\right)\frac{1}{k-1}+O(\theta^3),\\
&= \theta^2 P_{ab}P_{ac}\sum_{k=2}^{n-1}\left[\frac{H_k}{k}-\frac{1}{k^2}+\frac{1}{k}-\frac{2}{k+1}\right]+O(\theta^3),
\end{aligned}
$$

which simplifies to (30) using (15). Using (A.6):

$$\begin{aligned}
p(E_{2\mathcal{S}}^{(b,c)}) &= \sum_{n_a=1}^{n-1} p(n_a \boldsymbol{e}_a + (n-n_a)\boldsymbol{e}_c, E_{2\mathcal{S}}^{(b,c)}), \\
&= \theta^2 P_{ab} P_{bc} \sum_{n_a=1}^{n-1} \sum_{l_y=1}^{n_a} \binom{n_a - 1}{l_y - 1} \binom{n-1}{l_y}^{-1} \frac{1}{l_y(l_y+1)} + O(\theta^3), \\
&= \theta^2 P_{ab} P_{bc} \sum_{l_y=1}^{n-1} \sum_{n_a=l_y}^{n-1} \binom{n_a - 1}{l_y - 1} \binom{n-1}{l_y}^{-1} \frac{1}{l_y(l_y+1)} + O(\theta^3), \\
&= \theta^2 P_{ab} P_{bc} \sum_{l_y=1}^{n-1} \frac{1}{l_y(l_y+1)} + O(\theta^3),
\end{aligned}$$

which gives (31). Using (A.7):

$$\begin{aligned}
p(E_{2\mathcal{B}}^{(b,c)}) &= \sum_{n_b=1}^{n-1} p(n_b \boldsymbol{e}_b + (n-n_b)\boldsymbol{e}_c, E_{2\mathcal{B}}^{(b,c)}), \\
&= \theta^2 P_{ab} P_{ac} \sum_{n_b=1}^{n-1} \sum_{m=1}^{n_b} \binom{n_b - 1}{m - 1} \binom{n-1}{m}^{-1} \frac{1}{m^2(m+1)} + O(\theta^3), \\
&= \theta^2 P_{ab} P_{ac} \sum_{m=1}^{n-1} \sum_{n_b=m}^{n-1} \binom{n_b - 1}{m - 1} \binom{n-1}{m}^{-1} \frac{1}{m^2(m+1)} + O(\theta^3), \\
&= \theta^2 P_{ab} P_{ac} \sum_{m=1}^{n-1} \frac{1}{m^2(m+1)} + O(\theta^3),
\end{aligned}$$

which gives (32). To obtain $p(E_{2\mathcal{N}})$, $p(E_{2\mathcal{N}\mathcal{N}})$, $p(E_{2\mathcal{S}})$, and $p(E_{2\mathcal{B}})$ from each of these results we simply sum b and c over $1, \ldots, K$. □

Proof of Theorem 6.1. The given expressions are obtained immediately from Theorem 5.2 and the following observations:

$$\begin{aligned}
O_1 \cap E_2 &= \bigcup_b E_{2\mathcal{S}}^{(b,a)}, \\
O_{1\mathcal{S}} \cap E_2 &= \bigcup_b E_{2\mathcal{B}}^{(b,b)},
\end{aligned}$$

$$O_2 \cap E_2 = \left[\bigcup_b E_{2\mathcal{N}}^{(b,a)}\right] \cup \left[\bigcup_b E_{2\mathcal{N}\mathcal{N}}^{(b,b)}\right] \cup \left[\bigcup_b \bigcup_{c\neq a} E_{2\mathcal{S}}^{(b,c)}\right],$$
$$O_{2\mathcal{S}} \cap E_2 = \bigcup_b \bigcup_{c\neq b} E_{2\mathcal{B}}^{(b,c)},$$
$$O_3 \cap E_2 = \left[\bigcup_b \bigcup_{c\neq a} E_{2\mathcal{N}}^{(b,c)}\right] \cup \left[\bigcup_b \bigcup_{c\neq b} E_{2\mathcal{N}\mathcal{N}}^{(b,c)}\right].$$

Notice that each of these unions is over disjoint sets, so we can simply sum over the relevant probabilities. For example,

$$p(O_1, E_2) = \sum_{b=1}^{K} p(E_{2\mathcal{S}}^{(b,a)}) = \theta^2 \sum_{b=1}^{K} P_{ab}P_{ba}\left[1 - \frac{1}{n}\right] + O(\theta^3),$$

which equals (33). The others follow similarly. □

Proof of Theorem 6.2. For $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, the expression (38) for $p(\boldsymbol{n} \mid O_3)$ follows from

$$p(\boldsymbol{n} \mid O_3) = \frac{p(\boldsymbol{n}, O_3, E_2)}{p(O_3, E_2)} + O(\theta),$$

where $p(O_3, E_2)$ is given by (37), and $p(\boldsymbol{n}, O_3, E_2)$ is obtained from

$$p(\boldsymbol{n}, O_3, E_2) = p(\boldsymbol{n}, E_{2\mathcal{N}}^{(b,c)}) + p(\boldsymbol{n}, E_{2\mathcal{N}}^{(c,b)}) + p(\boldsymbol{n}, E_{2\mathcal{N}\mathcal{N}}^{(b,c)}) + p(\boldsymbol{n}, E_{2\mathcal{N}\mathcal{N}}^{(c,b)}),$$

with the right-hand side given by equations (18) and (20). □

Proof of Theorem 6.3. We use induction on what is sometimes referred to as the *sample complexity*, $n + l$. We show that, for a sample configuration $\boldsymbol{n}$ with l observed alleles including the ancestral allele:

$$p(\boldsymbol{n}, O_l) \propto \theta^{l-1} \prod_{k\in\Lambda} \frac{P_k}{n_k} + O(\theta^l), \tag{A.8}$$

the result following by dividing by $p(O_l) = O(\theta^{l-1})$ and letting $\theta \to 0$. We have already seen (A.8) to hold for $l = 2$ and $l = 3$, and all n [equations (17) and (40)], and we suppose inductively that it holds for all samples with complexity less than $n+l$. Let $\boldsymbol{n}$ be a suitable sample with complexity $n+l$.

Substituting the inductive hypothesis into the first (coalescence) term on the right-hand side of (3) and simplifying, we obtain

$$\frac{\theta^{l-1}}{n-1+\theta}\left(\prod_{k\in\Lambda}\frac{P_k}{n_k}\right)\left[n_a-1+\sum_{\substack{j:\,j\neq a,\\ n_j\geq 2}} n_j\right].$$

For the mutation term on the right of (3), the key here is to observe that contributions of $O(\theta^{l-1})$ come *only* from configurations $\boldsymbol{n}-\boldsymbol{e}_j+\boldsymbol{e}_i$ with $l-1$ observed alleles and including the ancestral allele; that is, from configurations in which the loss of one gamete with allele j and gain of one gamete with allele i reduces the number of observed mutant alleles in $\boldsymbol{n}$ by one. Thus, we must have had $n_j=1$, $j\neq a$, $n_i\geq 1$, and $i\neq j$, with any other possibilities requiring additional mutation events and thus contributing only higher order terms. The mutation term in (3) is therefore proportional to

$$\begin{aligned}&\frac{\theta}{n-1+\theta}\sum_{\substack{j:\,j\neq a,\\ n_j=1}}\sum_{\substack{i:\,i\neq j,\\ n_i\geq 1}} P_j\left(\frac{n_i+1}{n}\right)\theta^{l-2}\left(\prod_{k\in\Lambda\setminus\{j\}}\frac{P_k}{n_k+\delta_{ik}}\right)+O(\theta^l)\\ &=\frac{\theta^{l-1}}{n-1+\theta}\sum_{\substack{j:\,j\neq a,\\ n_j=1}}\left(\prod_{k\in\Lambda}\frac{P_k}{n_k}\right)+O(\theta^l).\end{aligned}$$

Summing contributions from both coalescence and mutation, (3) tells us

$$p(\boldsymbol{n})\propto\frac{\theta^{l-1}}{n-1+\theta}\left(\prod_{k\in\Lambda}\frac{P_k}{n_k}\right)(n-1)+O(\theta^l)\quad=\theta^{l-1}\left(\prod_{k\in\Lambda}\frac{P_k}{n_k}\right)+O(\theta^l),$$

as required. □

Proof of Theorem 7.1. The argument parallels that of Hobolth and Wiuf (2009), who obtained the corresponding result for two *nested* mutations. We first condition on the number k of lineages at the time of the younger mutation. Inspection of (A.5) yields

$$\begin{aligned}p(\boldsymbol{n},k,E^{(b,c)}_{2\mathcal{N}\mathcal{M}})=\frac{\theta^2P_{ab}P_{ac}}{n-1}\Bigg[&\binom{n_a+n_b-1}{k-2}\frac{1}{n_b}-\binom{n_a-1}{k-2}\frac{1}{n_b}\\ &-\binom{n_a+n_b-2}{k-3}\frac{1}{k}\Bigg]\binom{n-2}{k-2}^{-1}+O(\theta^3).\end{aligned}$$

Letting $\theta \to 0$, $n_b/n \to f_b$ and $n_c/n \to f_c$ while $n \to \infty$, we find

$$p(k \mid E^{(b,c)}_{2\mathcal{NN}}, \boldsymbol{f}) = \frac{1}{F}\left[\frac{(1-f_c)^{k-2}}{f_b} - \frac{(1-f_b-f_c)^{k-2}}{f_b} - \frac{k-2}{k}(1-f_c)^{k-3}\right], \tag{A.9}$$

with normalizing constant

$$\begin{aligned} F &= \sum_{k=3}^{\infty}\left[\frac{(1-f_c)^{k-2}}{f_b} - \frac{(1-f_b-f_c)^{k-2}}{f_b} - \frac{k-2}{k}(1-f_c)^{k-3}\right], \\ &= \frac{1}{f_c(f_b+f_c)} - \frac{2\ln f_c}{(1-f_c)^3} - \frac{1+f_c}{f_c(1-f_c)^2}. \end{aligned}$$

Under the standard, neutral coalescent model, the mean age of the younger mutation when it occurred during the time that there existed k ancestral lineages is $2/(k-1)$ (Hobolth and Wiuf, 2009). Hence

$$\mathbb{E}[A_c \mid E^{(b,c)}_{2\mathcal{NN}}, \boldsymbol{f}] = \sum_{k=3}^{\infty} \frac{2}{k-1} p(k \mid E^{(b,c)}_{2\mathcal{NN}}, \boldsymbol{f}).$$

Substituting in (A.9), summing over k and simplifying recovers (45). □

References

Achaz, G., 2009. Frequency spectrum neutrality tests: one for all and all for one. Genetics 183, 249–258.

Bustamante, C. D., Wakeley, J., Sawyer, S., Hartl, D. L., 2001. Directional selection and the site-frequency spectrum. Genetics 159, 1779–1788.

Desai, M. M., Plotkin, J. B., 2008. The polymorphism frequency spectrum of finitely many sites under selection. Genetics 180, 2175–2191.

Evans, S. N., Shvets, Y., Slatkin, M., 2007. Non-equilibrium theory of the allele frequency spectrum. Theor. Popul. Biol. 71, 109–119.

Fu, Y.-X., 1995. Statistical properties of segregating sites. Theor. Popul. Biol. 48, 172–197.

Griffiths, R. C., 2003. The frequency spectrum of a mutation, and its age, in a general diffusion model. Theor. Popul. Biol. 64, 241–251.

Griffiths, R. C., Tavaré, S., 1994. Simulating probability distributions in the coalescent. Theor. Popul. Biol. 46, 131–159.

Griffiths, R. C., Tavaré, S., 1998. The age of a mutation in a general coalescent tree. Stoch. Models 14, 273–195.

Griffiths, R. C., Tavaré, S., 2003. The genealogy of a neutral mutation. In: Green, P., Hjort, N., Richardson, S. (Eds.), Highly structured stochastic systems. Oxford University Press, pp. 393–412.

Hobolth, A., Wiuf, C., 2009. The genealogy, site frequency spectrum and ages of two nested mutant alleles. Theor. Popul. Biol. 75, 260–265.

Hodgkinson, A., Eyre-Walker, A., 2010. Human triallelic sites: evidence for a new mutational mechanism? Genetics 184, 233–241.

Johnson, P. L. F., Slatkin, M., 2006. Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res. 16, 1320–1327.

Kimura, M., Ohta, T., 1973. The age of a neutral mutation persisting in a finite population. Genetics 75, 199–212.

Kingman, J. F. C., 1982. The coalescent. Stoch. Proc. Appl. 13 (3), 235–248.

Polanski, A., Kimmel, M., 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165, 427–436.

Roman, S., 1993. The harmonic logarithms and the binomial formula. J. Comb. Theory A 63, 143–163.

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

Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10 (3), 512–526.

Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26, 119–164.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.

Watterson, G. A., 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7, 256–276.

Wiuf, C., Donnelly, P., 1999. Conditional genealogies and the age of a neutral mutant. Theor. Popul. Biol. 56, 183–201.

Wright, S., 1949. Adaptation and selection. In: Jepson, G. L., Mayr, E., Simpson, G. G. (Eds.), Genetics, Paleontology and Evolution. Princeton University Press, pp. 365–389.

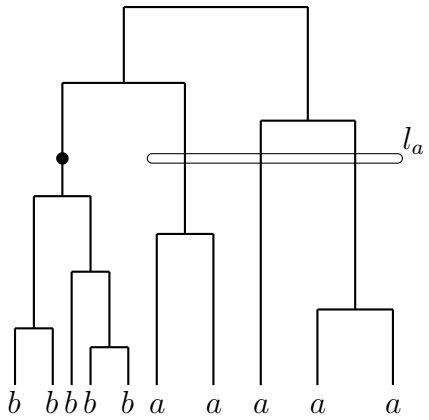

Figure 1: A coalescent tree with one mutation. The allele of each leaf is annotated. Also annotated is the variable l_a (here, $l_a = 3$), which determines the number of each type at the time of the mutation event.

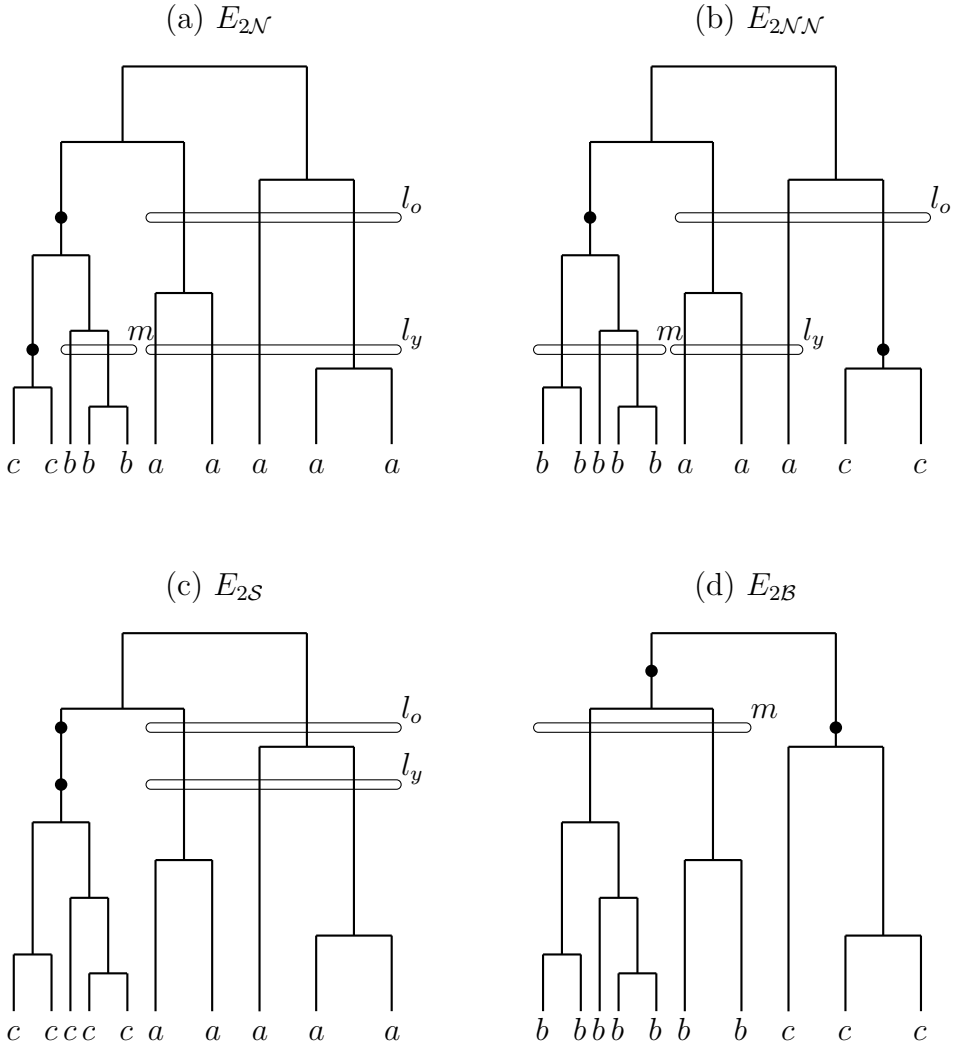

Figure 2: Coalescent trees with two mutations. (a) Two nested mutations. (b) Two nonnested mutations. (c) Two mutations on the same branch. (d) Two mutations on the basal branches. The allele of each leaf is annotated. Also annotated are variables determining the number of each type at the times of the mutation events; for example, in (a) we have $m = 2$, $l_y = 4$, and $l_o = 3$.

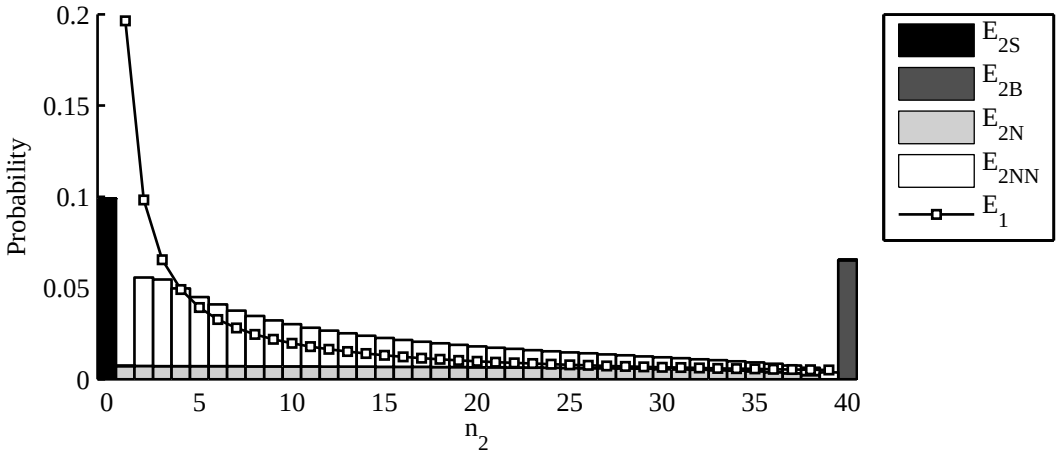

Figure 3: The sample frequency spectrum for a diallelic model with $n = 40$, conditional on one or two mutation events having occurred (line and stacked bars, respectively), as $\theta \to 0$. Monomorphic samples that are a result of two mutations have been included ($E_{2\mathcal{S}}$ and $E_{2\mathcal{B}}$); hence, to be directly comparable the plot for one mutation has been scaled down to sum to $p(E_{2\mathcal{N}} \mid E_2) + p(E_{2\mathcal{NN}} \mid E_2)$.

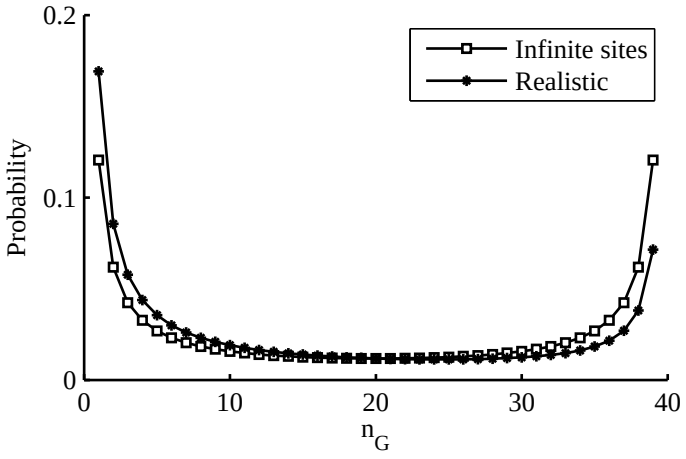

Figure 4: The sample frequency spectrum of a sample of size $n = 40$ under realistic assumptions about rates of mutation. The rate matrix P and a distribution over the ancestral allele are based on empirical data described in the main text [see equation (27)]. Also shown is the usual, infinite sites, frequency spectrum when the ancestral allele is unknown (i.e. the folded frequency spectrum, unfolded here for comparison).

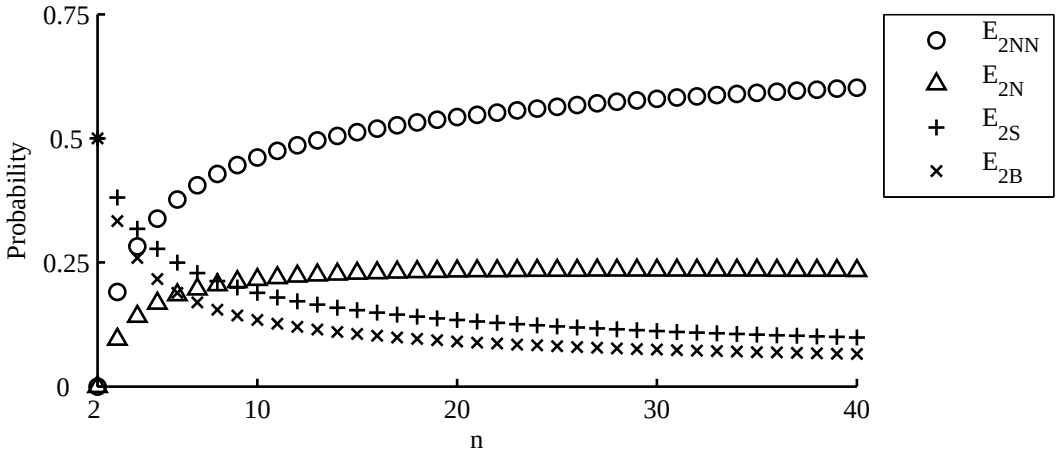

Figure 5: The probability of each of the four possible topologies conditional on two mutation events as $\theta \to 0$. Plots are for two nonnested mutations ($E_{2\mathcal{NN}}$), two nested mutations ($E_{2\mathcal{N}}$), two mutations on the same branch ($E_{2\mathcal{S}}$), and two mutations on the basal branches ($E_{2\mathcal{B}}$).

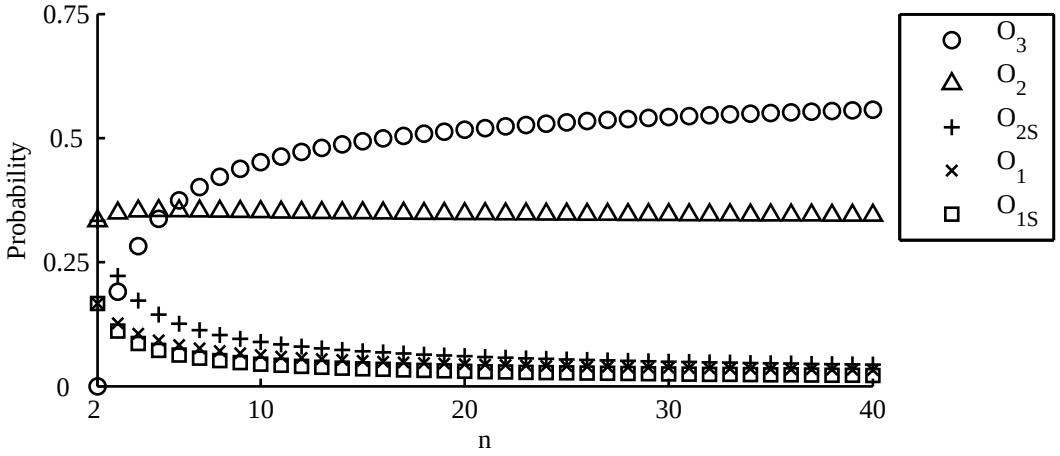

Figure 6: The probability of each possible observable outcome given two mutation events, as $\theta \to 0$: a triallelic polymorphism (O_3), a regular polymorphism with one allele ancestral and one mutant (O_2), a polymorphism in which both observed alleles are mutant ($O_{2\mathcal{S}}$), the entire sample is ancestral (O_1), and the entire sample has a mutant allele ($O_{1\mathcal{S}}$). Here we take a mutation model of four alleles, in which any mutation is to one of the other alleles with probability 1/3.

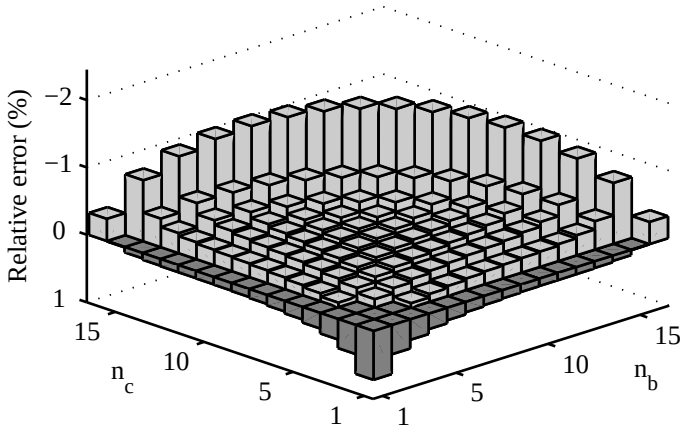

Figure 7: The (signed) relative error from assuming the sample frequency spectrum of a quadrallelic site is proportional to $(n_b n_c n_d)^{-1}$ [equation (48)]. Here, the sample size is $n = 20$ and the mutation parameter is $\theta = 0.01$. Shown are relative errors for a representative slice through the simplex of configurations, defined by fixing $n_d = 2$. Positive relative errors are shown in dark grey, negative relative errors in light grey.

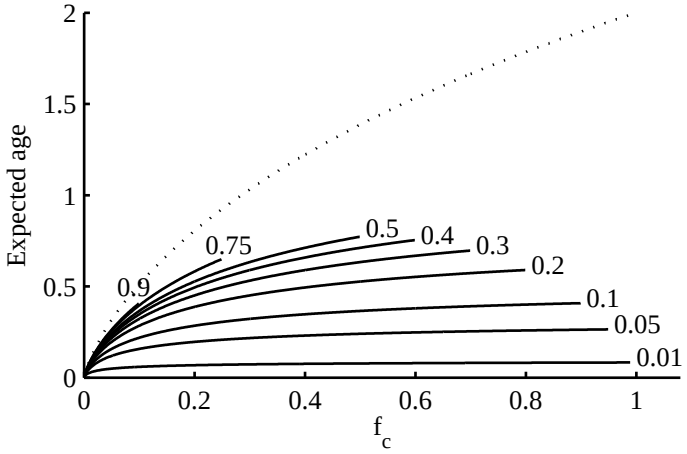

Figure 8: The expected age of the younger of two mutant alleles at a triallelic site [equation (47)]. The mutant alleles are at frequencies f_b (annotated) and f_c (x-axis). The expected age of a single mutant at a diallelic site and at frequency f_c is shown for comparison (dotted line).

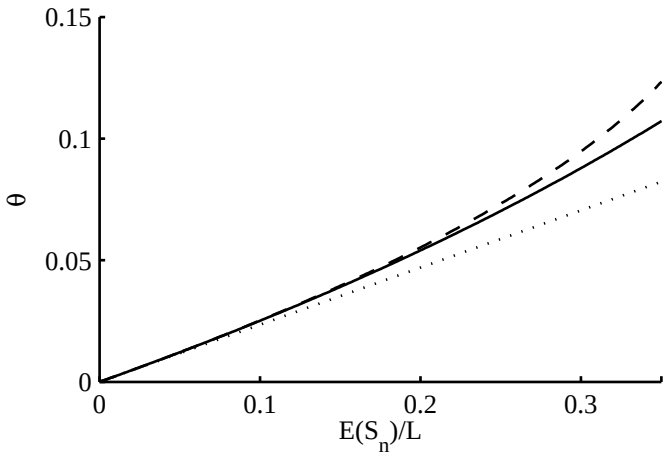

Figure 9: Estimating θ per site by the number s of segregating sites, out of L sites in total. The sample size is $n = 40$. Estimation of θ assumes at most one mutation event per site [dotted line, (49)], or at most two mutation events per site [dashed line, (52)]. The true relationship between θ and $\mathbb{E}(S_n)/L$ is plotted as a solid line.

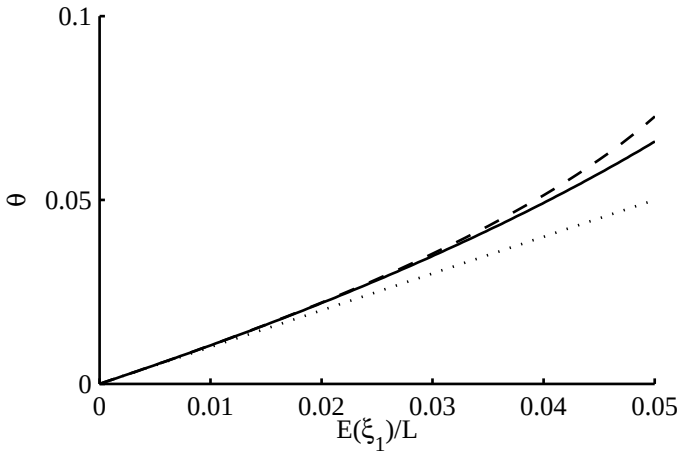

Figure 10: Estimating θ per site by the number ξ_1 of singleton segregating sites, out of L sites in total. The sample size is $n = 40$. Estimation of θ assumes at most one mutation event per site [dotted line, (50)], or at most two mutation events per site [dashed line, (53)]. The true relationship between θ and $\mathbb{E}(\xi_1)/L$ is plotted as a solid line.