Recurrent mutation in the ancestry of a rare variant

John Wakeley[1,*,†], Wai-Tong (Louis) Fan[2,3,†], Evan Koch[4,5], and Shamil Sunyaev[4,5]

[1] *Department of Organismic and Evolutionary Biology, Harvard University*
[2] *Department of Mathematics, Indiana University, Bloomington*
[3] *Center of Mathematical Sciences and Applications, Harvard University*
[4] *Department of Biomedical Informatics, Harvard Medical School*
[5] *Division of Genetics, Brigham and Women's Hospital, Harvard Medical School*
[†] *These authors contributed equally to this work.*
[*] *Corresponding author: wakeley@fas.harvard.edu*

August 18, 2022

Abstract

Recurrent mutation produces multiple copies of the same allele which may be co-segregating in a population. Yet most analyses of allele-frequency or site-frequency spectra assume that all observed copies of an allele trace back to a single mutation. We develop a sampling theory for the number of latent mutations in the ancestry of a rare variant, specifically a variant observed in relatively small count in a large sample. Our results follow from the statistical independence of low-count mutations, which we show to hold for the standard neutral coalescent or diffusion model of population genetics as well as for more general coalescent trees. For populations of constant size, these counts are given by the Ewens sampling formula. We develop a Poisson sampling model for populations of varying size, and illustrate it using new results for site-frequency spectra in an exponentially growing population. We apply our model to a large data set of human SNPs and use it to explain dramatic differences in site-frequency spectra across the range of mutation rates in the human genome.

Recurrent mutation has long been recognized as an important factor of evolution (Fisher, 1928; Haldane, 1933; Wright, 1938). This is emphasized by recent analyses of single-nucleotide polymorphism (SNP) frequencies and variation of mutation rates across the human genome (Aggarwala and Voight, 2016; Harpak et al., 2016; Seplyarskiy et al., 2021) describing how patterns of variation depend on the mutation rate, particularly for rare variants. By a rare variant we mean an allele, such as an alternate base at a SNP, which is observed a relatively small number of times in a large sample. Unless the mutation rate is very small, indistinguishable copies of the same allele may descend from multiple mutations. Here we present a sampling theory for the numbers and associated frequencies of these unobserved or latent mutations in the ancestry of a rare variant.

Humans are on the low end of polymorphism levels among species (Leffler et al., 2012). On average, multiple mutations should be rare. In the 1000 Genomes Project data, about 1 in 1300 sites differ when two (haploid) genomes are compared, and SNPs with more than two bases segregating comprise only about 0.3% of the total SNPs observed (The 1000 Genomes Project Consortium, 2015). But polymorphism rates vary by two or three orders of magnitude depending on local sequence context (Aggarwala and Voight, 2016; Harpak et al., 2016; Seplyarskiy et al., 2021). Recurrent mutation is an important phenomenon for fast-mutating sites. Evidence for this can be found in the haplotype structure surrounding rare mutations (Johnson and Voight, 2020) and in the distribution of their frequencies among sites in large samples (Harpak et al., 2016; Seplyarskiy et al., 2021).

Here we focus on the latter, in particular on the site-frequency spectrum (Tajima, 1989; Braverman et al., 1995; Fu, 1995). Deviations in site-frequency spectra compared to standard predictions may be due to selection (Bustamante et al., 2001; Achaz, 2009; Ferretti et al., 2017), changes in population size over time (Eldon et al., 2015; Liu and Fu, 2015; Gao and Keinan, 2016) or population structure (Gutenkunst et al., 2009; Städler et al., 2009; Kern and Hey, 2017). But they may also be due to multiple mutations, i.e. to violations of the infinite-sites model assumption that each polymorphism is due to a unique mutation (Fisher, 1930a; Kimura, 1969, 1971; Ewens, 1974; Watterson, 1975).

The standard site-frequency prediction, which holds for a well-mixed population of constant large size N and neutral mutation rate u at a locus, is that the number of SNPs where a variant is found in i copies in a sample of size n should be proportional to θ/i, where $\theta = 4Nu$ (Tajima, 1989; Fu, 1995). This dramatically underpredicts the abundance of rare variants in data from humans, which is largely due to our recent explosive population growth (Keinan and Clark, 2012; Gazave et al., 2014; Gao and Keinan, 2016), but the standard neutral model is a useful starting point for modeling recurrent mutation.

Jenkins and Song (2011) studied the occurrence of one or two mutations at a single site under the standard neutral coalescent model (Kingman, 1982; Hudson, 1983; Tajima, 1983). They showed that if two mutations occur and are non-nested (meaning that all descendants of both mutations can be observed) there will be a shift away from rare variants and toward common ones. An earlier work focusing on the nested case is Hobolth and Wiuf (2009). Bhaskar et al. (2012) used a similar approach as Jenkins and Song (2011) to obtain results for one, two or three mutations, up to leading order in the mutation parameter θ. Sargsyan (2006, 2015) considered two mutations occurring at two different sites, and Jenkins et al. (2014) assume that two mutations are distinguishable and yield a tri-allelic polymorphism. These latter works (Sargsyan, 2006, 2015; Jenkins et al., 2014) allowed for variable population size following the general coalescent approach of Griffiths and Tavaré (1998). None of these works considered rare variants in particular but their predictions, especially those for non-nested mutations (Jenkins and Song, 2011; Bhaskar et al., 2012) are helpful for understanding recurrent mutation.

Two recent large studies of human SNPs observed this predicted shift away from rare variants and toward common ones at fast-mutating sites. Harpak et al. (2016) surveyed about 8 million SNPs in a sample of nearly 61 thousand people in version 0.2 of the Exome Aggregation Consortium database (Lek et al., 2016) for which data were available from other primate species. Among these, about 93.3% of these were bi-allelic, 6.5% were tri-allelic and 0.2% were quad-allelic. Harpak et al. (2016) took the presence of identical segregating variants in different species, ranging from chimpanzees to baboons, as indicative of a higher mutation rate at a site. Consistent with the hypothesis of multiple latent mutations at fast-mutating sites, they found fewer rare variants at bi-allelic SNPs for which the minor allele was segregating in another species, and that this effect is stronger when the other species is closer to humans.

The work we present here builds upon the second of these studies. Seplyarskiy et al. (2021) looked at rare variants in two datasets, one containing about 292 million variants among nearly 43 thousand individuals in TOPMed freeze 5 (Taliun et al., 2021) and the other containing about 182 million variants among 15 thousand individuals in gnomAD version r2.0.2 (Karczewski et al., 2020). Variants were divided into 192 types: each of the 3 possible base substitutions at the middle site of all 64 possible trinucleotides. A classic example of a fast-mutating site in this context would be

ACG, which readily changes to ATG via a C to T transition at the CpG dinucleotide (Bird, 1980; Goldman, 1993). The main goals in Seplyarskiy et al. (2021) were to quantify how the rates of each kind of mutation vary across the genome and to partition this variation into distinct components correlated with different mutational processes.

Another aim, taken up in the Supplementary Materials of Seplyarskiy et al. (2021), was to correct for multiple mutations contributing to rare variants. Recurrent mutation was modeled as a multi-type Poisson process where mutations with lower sample counts occur independently at a locus to generate the appearance of higher count mutations (Desai and Plotkin, 2008). The expected counts in the absence of recurrence were taken from the site-frequency spectrum at slow-mutating sites. The loss of rare variants due to recurrent mutation at fast-mutating sites was quantified for sites with up to 70 copies of a rare variant. These were considered to have descended from up to 5 mutations. Slow-mutating sites, even with rates up to the genome average in humans, should conform fairly well to the infinite-sites assumption. Resampling from these as in Seplyarskiy et al. (2021) is a way of controlling for the myriad unknown factors affecting the site-frequency spectrum, including growth.

In this work, we present a sampling theory for latent mutations of rare variants at each given site-frequency count in a large sample. We describe a mathematical population genetic framework for the Poisson-resampling method in Seplyarskiy et al. (2021) and provide closed-form analytical expressions for several quantities of interest. We obtain new large-sample results for exponential growth and use these to illustrate the theory. We apply our results to a different subset of the gnomAD data than Seplyarskiy et al. (2021), synonymous variants observed in non-Finnish European individuals in v2.1.1, containing about 834 thousand variants at about 12.3 million sites among 57K individuals, presorted into 97 bins based on estimates of mutation rate by the method of Seplyarskiy et al. (2022, in prep.).

We develop and present these results in the next three sections. In Section 1, we begin with the standard neutral coalescent or diffusion model of population genetics (Ewens, 2004) and demonstrate a close connection between the Ewens sampling formula (Ewens, 1972) and distributions of latent mutations. In Section 2, we extend the results to populations which have changed in size, using the Poisson-sampling models of Watterson (1974a) and Arratia et al. (1992). In Section 3, we compare predictions for constant size to those for exponential growth and show how the new theory can be applied to understand the effects of recurrent mutation on counts of rare variants across the range of human per-site mutation rates.

1 Theory for constant-size large populations

In this section, we begin with a description of recurrent mutation via the well known predictions for allele frequencies in a population and in a sample at stationarity. We then use conditional ancestral processes to demonstrate independence of latent mutations of rare variants in a large sample, and show that their distribution is given by the Ewens sampling formula.

1.1 Stationary distributions and sampling probabilities

Consider a single locus with parent-independent mutation among K possible alleles in a population which obeys the Wright-Fisher diffusion (Fisher, 1930b; Wright, 1931; Ewens, 2004). Thus, the population is very large, well mixed, constant in size over time, and there is no selection. One unit of time in the diffusion process corresponds to $2N_e$ generations (N_e generations for haploid species) where N_e is the effective population size. Each gene copy or genetic lineage experiences mutations at rate $\theta/2$ and each mutation produces an allele of type $i \in (1, \ldots, K)$ with probability π_i, with $\sum_i \pi_i = 1$, independent of the allelic state of the parent. At stationarity, the joint distribution of the relative frequencies $x_1, \ldots, x_{K-1}$ of alleles is given by

$$\phi(x_1, \ldots, x_{K-1}) = \Gamma(\theta) \prod_{i=1}^{K} \frac{x_i^{\theta\pi_i - 1}}{\Gamma(\theta\pi_i)} \tag{1}$$

in which $\Gamma(\cdot)$ is the Gamma function, and where necessarily $x_K = 1 - \sum_{i<K} x_i$ (Wright, 1931, 1949).

Conditional on the population frequencies $(X_1, \ldots, X_K)$ the sample counts of alleles $(\mathcal{N}_1, \ldots, \mathcal{N}_K)$ are multinomially distributed. A sample of size n taken from the population contains $n_1, \ldots, n_{K-1}$ copies of alleles 1 though $K-1$, and necessarily $n_K = n - \sum_{i<K} n_i$ copies of allele K, with probability

$$\begin{aligned} p(n_1, \ldots, n_{K-1}; n) &\equiv \mathbb{P}\left[\mathcal{N}_1 = n_1, \ldots, \mathcal{N}_{K-1} = n_{K-1}; n\right] && (2) \\ &= \binom{n}{n_1 \cdots n_K} \mathbb{E}\left[X_1, \ldots, X_{K-1}\right] && (3) \\ &= \binom{n}{n_1 \cdots n_K} \left(\theta_{(n)}\right)^{-1} \prod_{i=1}^{K} (\theta\pi_i)_{(n_i)} && (4) \end{aligned}$$

for $n_i \in (0, 1, \ldots, n)$ constrained by $\sum_i n_i = n$ and where $k_{(r)}$ denotes the Pochhammer function or rising factorial $k(k+1)\cdots(k+r-1)$ with $k_{(0)} = 1$. The shorthand defined in (2) is used extensively in what follows.

In applications to DNA, $K = 4$ and a sample at a given site would contain counts n_1, n_2, n_3, n_4 of each of the four nucleotides. The assumption of parent-independent mutation which leads to the relatively simple expressions (1) and (4) is unrealistic for DNA, but its results are useful in the case of rare variants in very large samples. In this case, it is likely that the common variant, allele 4 say, represents the ancestral state of the entire sample and that rare variants (alleles 1, 2 and 3) are due to recent mutations from the common variant. Then the mutation parameter $\theta\pi_i$ for $i \in (1, 2, 3)$ captures the production of type-i rare alleles in a specific ancestral background (allele 4).

An instructive special case is $K = 2$, where we have

$$\phi(x) = \frac{\Gamma(\theta)}{\Gamma(\theta\pi_1)\Gamma(\theta\pi_2)} x^{\theta\pi_1 - 1}(1-x)^{\theta\pi_2 - 1} \tag{5}$$

for the stationary distribution of the frequency of type 1 in the population Wright (1931), and

$$p(n_1; n) = \binom{n}{n_1} \frac{(\theta\pi_1)_{(n_1)}(\theta\pi_2)_{(n-n_1)}}{\theta_{(n)}} \tag{6}$$

for the sampling probability, i.e. that a sample of size n contains n_1 copies of allele 1 and $n_2 = n - n_1$ copies of allele 2. Any two-allele mutation model can be described as a parent-independent model, but this is not so in general for $K > 2$.

Figure 1 shows how the sample frequency distribution $p(n_1; n)$ in (6) depends on the mutation rate for a pair of alleles which differ by an order of magnitude in mutation rate. Three value of θ are shown (small, blue; middle, orange; large, red) with the small value chosen so that the mutation rate for allele 2 ($\theta\pi_2$) is equal to the human average of about 1/1300 and the mutation rate for allele 1 ($\theta\pi_1$) is ten times that. When θ is small, the distribution is U-shaped and nearly symmetric,

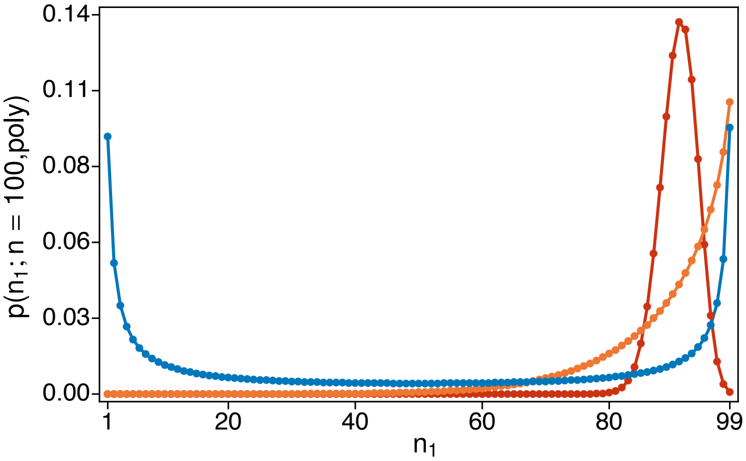


Figure 1: Sample frequency distribution $p(n_1; n)$ for $n = 100$, with $\pi_1 = 10\pi_2$ and three values of θ (smallest in blue, medium in orange, largest in red). The smallest θ was chosen so that $\theta\pi_2 = 1/1300 \sim 0.00077$, i.e. around the human average. The value of θ increases one-thousand fold from smallest to medium and again from medium to largest. In all three cases, the probabilities are normalized to sum to one, i.e. conditioned on the sample being polymorphic ($1 \leq n_1 \leq 99$).

given that the sample is polymorphic. When θ is around one, the distribution becomes J-shaped (or L-shaped if $\pi_1 < \pi_2$). When θ is large, the distribution has a peak around π_1. Graphs of $\phi(x)$ (not shown) display these same shapes, and $p(n_1; n)$ will be very close to $\phi(x)dx$ when n is large.

1.1.1 Relationship to infinite-sites frequency spectra

We use θ for the per-site mutation parameter. In a collection of L total sites at which (6) holds, the finite-sites version of the site-frequency spectrum (i.e. the expected number of sites with n_1 copies of allele 1 and n_2 copies of allele 2) is given the product $Lp(n_1; n)$. Infinite-sites mutation models may be obtained as limits of finite-sites models as L tends to infinity with the total mutation parameter $L\theta$ remaining finite. So when θ is small, we expect finite-sites results to be close to the usual (infinite-sites) predictions from the diffusion model (Ewens, 1979, 2004) or the coalescent model (Fu, 1995). Finite-sites models distinguish between kinds of mutations, subject to different mutation pressures, whereas infinite-sites models implicitly treat all mutations the same.

From Ewens (1979) equation (8.18) or Ewens (2004) equation (9.18)—see also Wright (1938) equation (16)—the expected number of sites segregating in the population with frequencies between x and $x + dx$ under the infinite-sites model is proportional to $1/x$. For comparison to (5) we may write

$$\phi_{ISM}(x) \propto \frac{\theta\pi_1}{x} \tag{7}$$

for a single site (θ small) approximately under the standard infinite-sites mutation model. For comparison with (6), we have

$$p_{ISM}(n_1; n) \propto \frac{\theta\pi_1}{n_1} \tag{8}$$

for the approximate single-site probability that there are n_1 type-1 alleles in a sample of size n. Equation (8) has the same form as the usual infinite-sites site-frequency spectrum (Fu, 1995) but here it is for a specific mutant (allele 1) with a specific ancestral type (allele 2 in the two-allele model).

From (5) and (6) with θ small we have

$$\phi(x) = \pi_2 \frac{\theta\pi_1}{x} + \pi_1 \frac{\theta\pi_2}{1-x} + O\left(\theta^2\right) \tag{9}$$

and

$$p(n_1; n) = \pi_2 \frac{\theta\pi_1}{n_1} + \pi_1 \frac{\theta\pi_2}{n_2} + O\left(\theta^2\right) \tag{10}$$

for $n_1 \in (1, \ldots, n-1)$. The diffusion result (5) does not admit atoms of probability at $x = 0$ or $x = 1$—see section 10.7 of Ewens (2004) for discussion—but we can interpret (9) intuitively as follows. If θ is close to zero, most of the time the population will be fixed, containing only allele 1 with probability π_1 and only allele 2 with probability π_2. Mutants of type 2 and type 1 are introduced with rates $\theta\pi_2$ and $\theta\pi_1$ in these two backgrounds, respectively. Then the leading terms in (9) represent a mixture of two infinite-sites models like (7) with the constants of proportionality specified. Equation (10) has an identical interpretation, as a mixture of two infinite-sites site-frequency spectra.

Although no closed-form expression like (1) is available except under parent-independent mutation, Burden and Tang (2016, 2017) have shown that the stationary densities for pairs of alleles under general mutation models take forms identical to (9) when θ is small; see equation (21) in Burden and Tang (2017). Similarly from a coalescent analysis of general K-alleles mutation, Bhaskar et al. (2012) obtained leading order terms for sampling probabilities with forms identical to (10) when θ is small and samples contain just two alleles. For $K = 2$, the result from Theorem 1 of Bhaskar et al. (2012) is identical to (10).

1.2 Mutation and the frequencies of rare sample variants

Our goal here is to understand how the frequency spectra of rare variants depend on θ and on the number of mutation events in the ancestry of the sample under the standard neutral coalescent or diffusion model of population genetics which assumes constant population size (Ewens, 2004). We first describe an ancestral process for the sample, then focus on rare variants in a large sample to obtain predictions about latent mutations.

1.2.1 A conditional ancestral process for rare variants

Here we focus on ordered samples because the calculations are more intuitively related to the familiar rates of events in the ancestral coalescent process. The results do not depend on the order and so apply equally to ordered and unordered samples. Using the subscript "o" for ordered and writing $p_{\text{o}}(n_1, \ldots, n_K)$ in place of $p_{\text{o}}(n_1, \ldots, n_{K-1}; n)$ to facilitate the calculations, we have

$$p_{\text{o}}(n_1, \ldots, n_K) = \left(\theta_{(n)}\right)^{-1} \prod_{i=1}^{K} (\theta\pi_i)_{(n_i)} \tag{11}$$

which differs from the sampling probability in (3) and (4) only by the multinomial coefficient, or the number of ways a sample containing allele counts $n_1, \ldots, n_K$ can be ordered.

Equation (11) is suggestive, as are (4) and (6), that the sampling structure of the n_i copies of allele i may be related to the Ewens sampling formula (Ewens, 1972). Specifically, from the fact that

$$(\theta\pi_i)_{(n_i)} = \sum_{k_i=1}^{n_i} S_{n_i}^{(k_i)} (\theta\pi_i)^{k_i}, \tag{12}$$

where $S_{n_i}^{(k_i)}$ is an (unsigned) Stirling number of the first kind, we might guess that there is a latent variable k_i which is the number of mutations giving rise to the n_i copies of allele i. As in the usual application of the Ewens sampling formula, in contrast to the total possible number of type-i mutations in the ancestry of the sample, these latent mutations are just those $k_i \in (1, \dots, n_i)$ most recent ones which produced the observed alleles.

That is, based on (11) and (12), we suppose that the joint probability of the sample counts $n_1, \dots, n_K$ and their numbers of latent mutations $k_1, \dots, k_K$ is given by

$$p_{\mathrm{o}}(k_1, \dots, k_K, n_1, \dots, n_K) = \left(\theta_{(n)}\right)^{-1} \prod_{i=1}^{K} S_{n_i}^{(k_i)} (\theta\pi_i)^{k_i}, \tag{13}$$

and therefore that the probability of $k_1, \dots, k_K$ conditional on $n_1, \dots, n_K$ is given by

$$p(k_1, \dots, k_K | n_1, \dots, n_K) = \prod_{i=1}^{K} \frac{S_{n_i}^{(k_i)} (\theta\pi_i)^{k_i}}{(\theta\pi_i)_{(n_i)}} \tag{14}$$

which applies to both ordered and unordered samples.

It is straightforward to check that (13) satisfies the corresponding recursive equation for the sampling probability, obtained using (17) in (15c) below and keeping track of latent mutations (not shown). It can also be obtained by the approach of Donnelly and Tavaré (1987), which begins with the Ewens sampling formula, labels mutations with allelic types 1 to K randomly (with probabilities π_1 to π_K in our notation) then retrieves the expected K-allele sampling probabilities.

Thus the number of latent mutations of an allele conditional on its sample count follows the Ewens sampling formula. But reckoned in this way under parent-independent mutation some of the latent mutations in (13) and (14) are ‘empty’ (Baake and Bialowons, 2008). They do not change the allelic type. These are a modeling artefact which must be dealt with not only in parent-independent models but in general mutation models as well the way these are typically implemented (Jenkins and Song, 2011; Bhaskar et al., 2012; Jenkins et al., 2014; Burden and Tang, 2017; Burden and Griffiths, 2019). Empty mutations have no empirical significance. Here we show that they almost never occur in the ancestry of rare variants in large samples.

We make use of the ancestral-process approach developed by Griffiths and Tavaré (1994a,b) based on recursive equations for sampling probabilities. For the K-allele model we have

$$\left(n\frac{\theta}{2} + \binom{n}{2}\right) p_{\mathrm{o}}(n_1, \dots, n_K) = \tag{15a}$$

$$\sum_{i=1}^{K} n_i \frac{\theta\pi_i}{2} \sum_{j=1}^{K} p_{\mathrm{o}}(\dots, n_i - 1, \dots, n_j + 1, \dots) \tag{15b}$$

$$+ \sum_{i=1}^{K} \binom{n_i}{2} p_{\mathrm{o}}(\dots, n_i - 1, \dots) \tag{15c}$$

with boundary conditions $p_{\mathrm{o}}(0, \dots, n_i = 1, \dots, 0) = \pi_i$ for $i \in (1, \dots, K)$. This is a recursion back into the ancestry of the sample, in which (15b) and (15c) include all events which could have produced the sample, and the probabilities of ancestral types required in each case to produce the sample. It can be derived either a using coalescent approach (De Iorio and Griffiths, 2004) or a diffusion approach (Burden and Griffiths, 2019). We have kept the common factors of 1/2 in

(15a), (15b) and (15c) to emphasize that the ancestral process occurs on the diffusion or coalescent time scale, with mutations happening at rate $\theta/2$ on each ancestral lineage and coalescent events happening at rate 1 for each pair of ancestral lineages.

Initially these ancestral processes were used to compute otherwise intractable likelihoods analytically for small samples or by simulation for large samples (Griffiths and Tavaré, 1994a,b). They were subsequently used to describe the joint sampling of gene genealogies and associated allelic types under selection (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997), and, in cases were the sampling probabilities are known, to describe conditional ancestral processes of samples given their allelic types (Slade, 2000a,b; Fearnhead, 2002; Stephens and Donnelly, 2003; Baake and Bialowons, 2008). We take the latter approach to obtain a large-n approximation for the conditional ancestral processes of mutation and coalescence for samples containing mostly one type. We set K to be the common allele, so we will have $n_K \gg n_1, \ldots, n_{K-1}$ and $n_K \sim n$.

Equations (15b) and (15c) include, respectively, all possible mutation events and all possible coalescent events in the ancestral process which have a non-zero chance of producing the sample $(n_1, \ldots, n_K)$. Note that the terms with $i = j$ in (15b) are the 'empty' mutations of Baake and Bialowons (2008).

Following Slade (2000b), the conditional ancestral process remains in state $(n_1, \ldots, n_K)$ for an exponentially distributed time, i.e. leaving that state at rate $n(\theta + n - 1)/2$, then jumps to each possible ancestral state with probabilities proportional to the terms in (15b) and (15c). Dividing through by the left-hand side, (15a), we have

$$1 = \sum_{i=1}^{K} \frac{n_i \theta \pi_i}{n(\theta + n - 1)} \sum_{j=1}^{K} \frac{p_\mathrm{o}(\ldots, n_i - 1, \ldots, n_j + 1, \ldots)}{p_\mathrm{o}(n_1, \ldots, n_K)} \tag{16a}$$

$$+ \sum_{i=1}^{K} \frac{n_i(n_i - 1) p_\mathrm{o}(\ldots, n_i - 1, \ldots)}{n(\theta + n - 1) p_\mathrm{o}(n_1, \ldots, n_K)} \tag{16b}$$

for the total probability of these events given that an event occurs in the ancestral process. These are, in (16a), mutations on lineages ancestral to alleles of type i which have alleles of type j as ancestors and, in (16b), coalescent events between lineages ancestral to alleles of type i.

Only the numbers ancestral lineages are tracked in (16a) and (16b). The full ancestry, or gene genealogy, can be modeled using exchangeability within each allelic type. That is, each of the $n_i(n_i - 1)/2$ pairs is equally likely to be involved in a type-i coalescent event and each of the n_i lineages is equally likely to be the one involved in a type-i mutation event.

Depending on what quantities or aspects of the ancestry are of interest, (16a) and (16b) may be augmented, simplified or otherwise rearranged. Here we follow Fearnhead (2002) and Baake and Bialowons (2008), in removing some lineages from the ancestral process once they have experienced a mutation. This is captured by the identity

$$p_\mathrm{o}(\ldots, n_i - 1, \ldots) = \sum_{j=1}^{K} p_\mathrm{o}(\ldots, n_i - 1, \ldots, n_j + 1, \ldots) \tag{17}$$

which can be used as needed in (16a). Our aim here is to model mutation and coalescence in the ancestry of the rare alleles with counts n_1 through n_{K-1} in the sample. So we follow the ancestry of the n_K common alleles only insofar as this affects the ancestries of the rare alleles. We use (17) to justify removing ancestral type-K lineages whenever they mutate, and we lump these events with coalescent events because their overall effect is the same ($n_K \to n_K - 1$). Additionally, we distinguish two kinds of mutation events among the rare alleles: ones in which the ancestral allele was the common allele K and ones in which it was a rare allele $j \in (1, \dots, K-1)$.

Making these changes, and using (11) to simplify the ratios of sampling probabilities, we have

$$\begin{aligned}
1 = &\sum_{i=1}^{K-1}\sum_{j=1}^{K-1} \frac{n_i\theta\pi_i(\theta\pi_j + n_j - \delta_{ij})}{n(\theta+n-1)(\theta\pi_i + n_i - 1)} && \text{(18a)}\\
&+ \sum_{i=1}^{K-1} \frac{n_i\theta\pi_i(\theta\pi_K + n_K)}{n(\theta+n-1)(\theta\pi_i + n_i - 1)} && \text{(18b)}\\
&+ \sum_{i=1}^{K-1} \frac{n_i(n_i-1)}{n(\theta\pi_i + n_i - 1)} && \text{(18c)}\\
&+ \frac{n_K}{n} && \text{(18d)}
\end{aligned}$$

in which we have used Kronecker's delta to accommodate empty mutations, $i = j$ in (18a). Recall that $n = \sum_i n_i$ which will be $O(n_K)$ when n_K becomes large for given n_1 though n_{K-1}. Equation (18a) contains the probabilities of all mutations on rare-allele lineages which have rare-allele ancestors. These probabilities are $O(1/n_K^2)$ when n_K is large. Equation (18b) contains the probabilities of all mutation events on rare-allele lineages which have common-allele ancestors. These are $O(1/n_K)$ when n_K is large. Equation (18c) contains the probabilities of all coalescent events between rare-allele lineages of the same type, similarly $O(1/n_K)$. Finally, (18d) gives the probability of mutation or coalescence among the common-allele lineages, which is $O(1)$.

Keeping only up to the $O(1/n_K)$ terms gives an approximate, large-n_K ancestral process with total rate $n(\theta+n-1)/2 \approx n_K^2/2$ and jumps, for $i \in (1, \dots, K-1)$, from state $(n_1, \dots, n_K)$ to state

$$(\dots, n_i - 1, \dots, n_K + 1) \quad \text{w/prob.} \quad \frac{n_i}{n_K}\frac{\theta\pi_i}{\theta\pi_i + n_i - 1}, \tag{19}$$

to state

$$(\dots, n_i - 1, \dots) \quad \text{w/prob.} \quad \frac{n_i}{n_K}\frac{n_i - 1}{\theta\pi_i + n_i - 1}, \tag{20}$$

or to state

$$(n_1, \dots, n_{K-1}, n_K - 1) \quad \text{w/prob.} \quad 1 - \frac{\sum_{i<K} n_i}{n_K}. \tag{21}$$

This process is dominated by (21), that is by events on lineages ancestral to the common allele K, which decrease the number of these but leave the counts of rare-allele lineages unchanged. Although we are not tracing the details of common-allele ancestry, we note that the overwhelming majority of these events will be coalescent events, since their rate is approximately equal to the total rate $\sim n_K^2/2$. The next most frequent will be empty mutation events at rate $O(n_K)$, followed by common-allele mutation events with rare-allele ancestors at rate $O(1)$.

When one of the rarer events occurs in the ancestral process, it involves allele i with probability n_i/n_K, then is either a mutation event from a common allele as in (19) or a coalescent event as in (20). For each allele $i \in (1, \dots, K-1)$ which is observed at least once in the sample, there will be exactly n_i such events. Again, the empty mutation events captured in (18a) are negligible for large n_K. Note that, the relative probabilities of mutation versus coalescence in (19) and (20) for each allele are identical to the standard ones from coalescent theory (Kingman, 1982), only here with $\theta\pi_i$ in place of the usual θ. It follows that both the number of latent mutations which produced the n_i copies and the counts of each mutation's descendants among the n_i copies are given by the Ewens sampling formula (Ewens, 1972; Kingman, 1982; Arratia et al., 1992, 2016).

The events involving the common allele in (21) occur very quickly. But since only a fixed number of events involving rare alleles are required to resolve the ancestry of latent mutation and coalescence, the approximation remains accurate until all the rare-allele events have happened, if n_K is large enough. In Appendix section A.1, we study the joint distribution of the times of events among the rare alleles and the numbers of common-allele ancestors when these rare-allele events occur. Focusing on the case of two alleles for simplicity, if $\mathcal{T}_i$ is the time back to the ith event involving the rare allele 1, we have

$$\mathbb{E}\left[\mathcal{T}_1\right] \approx \begin{cases} \frac{2\log(n_2)}{n_2} & \text{if} \quad n_1 = 1 \\ \frac{2}{n_2(n_1-1)} & \text{if} \quad n_1 > 1 \end{cases} \tag{22}$$

which in either case tends to zero as n_2 tends to infinity. Further, if $\mathcal{N}_2(\mathcal{T}_i)$ is the random number of type-2 ancestral lineages left at the ith event involving the rare allele 1, we have

$$\mathbb{E}\left[\mathcal{N}_2(\mathcal{T}_i)\right] \approx n_2 \frac{n_1 - i + 1}{n_1 + 1} \tag{23}$$

suggesting that, despite the rapid decrease of common-variant lineages, the approximation can hold until the entire ancestry of latent mutation and coalescence is resolved.

Even for the largest rare-variant site-frequency count considered in Seplyarskiy et al. (2021), there will still be $>$ 1200 common-variant lineages left on average at $\mathcal{T}_{70}$ for the TOPMed data ($n_2 \sim 86000$) and $>$ 400 left for the gnomAD data ($n_2 \sim 30000$). In Section 3.2, we consider site-frequency counts up to 40 for synonymous exonic sites in gnomAD with many fewer SNPs but a larger sample size ($n_2 \sim 114000$) and in this case there should be about 2780 common-variant lineages left at $\mathcal{T}_{40}$ when the entire ancestry of latent mutation and coalescence among the rare variants is resolved.

Thus, rare alleles in a large sample will quickly coalesce and mutate. Their ancestors will be common alleles. If $k_i \in (1, \ldots, n_i)$ is the number of mutations in the ancestry of allele $i \in (1, \ldots, K-1)$, then from the rates of mutation and coalescence in (19) and (20) we have

$$p(k_1, \ldots, k_{K-1} | n_1, \ldots, n_{K-1}; n \text{ large}) \approx \prod_{i=1}^{K-1} \frac{S_{n_i}^{(k_i)} (\theta\pi_i)^{k_i}}{(\theta\pi_i)_{(n_i)}} \tag{24}$$

which is the product of independent Ewens distributions.

1.2.2 Latent mutations and sample counts of rare alleles

Here we focus on the case in which a single type of mutation or allele is observed against a background of a given common allele, as in recent empirical studies (Harpak et al., 2016; Seplyarskiy et al., 2021). Our goal is to understand how counts of these mutant alleles depend on the number of latent mutations and on the mutation rate. As before, allele 1 is the focal rare allele and allele K is the common allele.

First, from (24) we have

$$p(k_1 | n_1; n \text{ large}) \approx \frac{S_{n_1}^{(k_1)} (\theta\pi_1)^{k_1}}{(\theta\pi_1)_{(n_1)}} \tag{25}$$

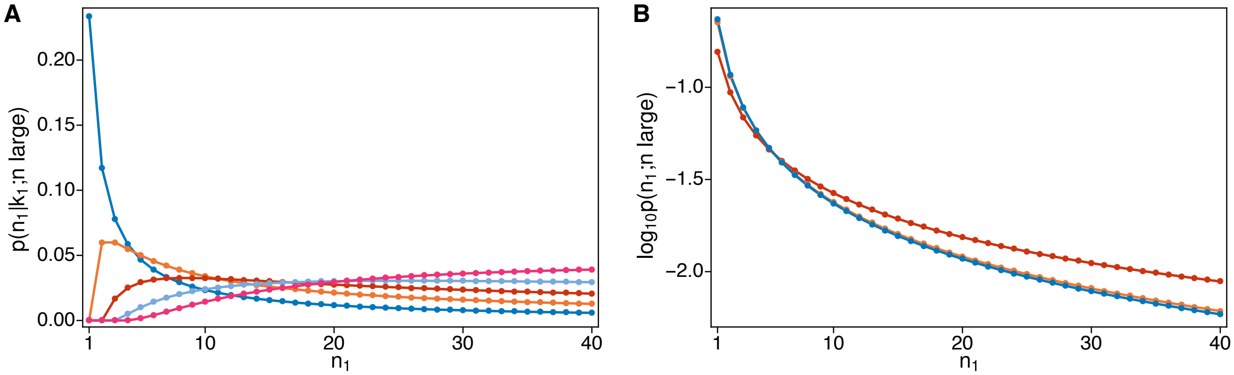


Figure 2: Panel A shows the probability of observing n_1 copies of allele 1 in a large sample given these are produced by $k_1 = 1, 2, 3, 4, 5$ mutations. Panel B shows the $\log_{10}$-probability of observing n_1 copies of allele 1 in a large sample for three different values of $\theta\pi_1$: 0.002, 0.02 and 0.2. Probabilities in both panels are normalized to sum to one for $n_1 \in (1, 2, \ldots, 40)$.

which sums to one for $k_1 \in (1, \ldots, n_1)$. To understand the effects of mutation on the number of rare alleles, we apply Stirling's formula as in 6.1.47 of Abramowitz and Stegun (1964) or equation (1) in Tricomi and Erdélyi (1951) to show that the main dependence of $p(n_1; n)$ on n_1 and $\theta\pi_1$ is captured by

$$p(n_1; n \text{ large}) \propto \frac{(\theta\pi_1)_{(n_1)}}{n_1!}. \tag{26}$$

Using this together with (25) we have

$$p(n_1|k_1; n \text{ large}) \propto \frac{S_{n_1}^{(k_1)}}{n_1!} \tag{27}$$

as the approximate dependence of the rare-allele count n_1 on the number of latent mutations k_1. Following the logic of Section 1.1.1, we can use (26) and (27) to understand how the site-frequency counts of a rare allele depend on the rate of production of the allele and on the number of latent mutations contributing to those counts.

The proportional relations (26) and (27) are sufficient for this if we adopt the usual convention of normalizing site-frequency counts to sum to one. Figure 2A shows how the site-frequency spectrum of a rare variant in a large sample depends on the number of mutations which produced the observed copies of the variant. When all copies descend from a single mutation ($k_1 = 1$), the usual predictions from the infinite-sites model hold. Thus, putting $S_{n_1}^{(1)} = (n_1 - 1)!$ in (27) we have

$$p(n_1|k_1 = 1; n \text{ large}) \propto \frac{1}{n_1}.$$

The total number of such sites will depend on $\theta\pi_1$, and in general on the factor $(\theta\pi_1)^{k_1}$ in (25) for larger numbers of latent mutations. But conditional on k_1, the normalized site-frequency counts for a rare variant do not depend on θ, at least to leading order in the sample size $n \sim n_K$. Further, if there are $k_1 > 1$ mutations in the ancestry of the rare variant, then n_1 cannot be less than k_1. This is illustrated in Figure 2A for $k_1 = 2$ to $k_1 = 5$ latent mutations. A key effect of recurrent mutation is to give relatively less weight to low site-frequency counts, as found previously by Jenkins and Song (2011).

Using (25) and (26) the joint distribution of n_1 and k_1 obeys

$$p(n_1, k_1; n \text{ large}) \propto \frac{S_{n_1}^{(k_1)} (\theta\pi_1)^{k_1}}{n_1!} \tag{28}$$

which can be compared to the results of Jenkins and Song (2011). With fixed n_1 and large n_K in our model, all mutations in the ancestry of the rare variant will be non-nested mutations; note this also follows from (18) in Jenkins and Song (2011). In addition, we have shown that the higher-order terms in θ, i.e. beyond $(\theta\pi_K)^{k_1}$, can be neglected when n_K is large. Adapting the notation of Jenkins and Song (2011) in which $E_{2\mathcal{N},\mathcal{N}}^{(1,1)}$ is the event that the n_1 copies of allele 1 are due to two non-nested mutations, both from allele K to allele 1, their (21) becomes

$$p\left(n_1, n_K, E_{2\mathcal{N},\mathcal{N}}^{(1,1)}\right) \approx \theta^2 \pi_1^2 \frac{S_{n_1}^{(2)}}{n_1!}$$

for large n_K (and small θ), which is identical to (28) if $k_1 = 2$.

Numerical computations (not shown) using the unnumbered equation below (10) in Jenkins and Song (2011), which holds for any θ, reproduce the case of $k_1 = 2$ shown in Figure 2A when n_K is large. This is evident in Figure 3 of Jenkins and Song (2011) for the quantity $E_{2\mathcal{N}\mathcal{N}}$. These computations are difficult for samples beyond the hundreds. Our results for $k_1 = 3$ could potentially also be compared to the $O(\theta^3)$ results of Bhaskar et al. (2012) using their Theorem 3 and summing appropriately.

Figure 2B shows how the site-frequency counts of the rare variant depend on the mutation parameter of that variant, $\theta\pi_1$. Although Figure 2A shows a dramatic effect of k_1 on the site-frequency counts, Figure 2B suggests that large values of k_1 are unlikely. This is evident from (25) and (28) in that each additional mutation results in an additional factor of $\theta\pi_1$. Note that the smallest value of $\theta\pi_1$ in Figure 2B is already more than twice the human average. From (26), we have

$$p(n_1; n \text{ large}, \theta \text{ small}) \propto \frac{\theta\pi_1}{n_1}$$

which is consistent with (10) in the case where allele 1 is rare in a large sample. Thus, when $\theta\pi_1$ is small (0.002 and 0.02 in Figure 2B) the site-frequency spectrum under recurrent mutation is very close to the standard infinite-sites model predictions. When $\theta\pi_1$ is large (0.2 in Figure 2B) the site-frequency spectrum under recurrent mutation is noticably different, with a dearth of low-frequency variants and corresponding excesses at higher frequencies. Figure 2B plots site frequencies on a log scale to better illustrate differences, especially at higher frequencies.

2 Theory for nonconstant populations

Here we extend our analysis to populations which deviate from the standard neutral site-frequency predictions. We have in mind populations which have changed in size, although other applications may be possible. Now gene genealogies are the general coalescent trees of Griffiths and Tavaré (1998), which have same the branching structure of standard coalescent trees but may have different distributions of coalescence times.

Equation (25) suggests another way to model both the number of copies (n_1) of a variant of interest and the corresponding count of latent mutations (k_1) when the variant is rare in a large sample. Arratia et al. (1992) proved that when the sample size tends to infinity, the numbers of alleles in small counts $1, 2, \ldots, i$ in the Ewens distribution converge to independent Poisson random variables with expected values $\theta, \theta/2, \ldots, \theta/i$. Note that θ/i is the usual expected site-frequency count of mutants in i copies in the sample under the standard neutral model of a large constant-size population. A seminal result of Watterson (1974a) is that the numbers and counts of mutations in a sample from such a multi-type Poisson distribution conform to the Ewens sampling formula when conditioned on their total size. So we may interpret (25) and other findings in the previous section within this independent-Poissons sampling framework.

This is exactly the approach in the Supplementary Materials of Seplyarskiy et al. (2021). Again, human SNP data strongly reject the standard neutral model with site-frequencies $\propto 1/i$, owing largely to the great excess of singletons and other rare variants due to our recent growth (Keinan and Clark, 2012; Gazave et al., 2014). So we replace $1/i$ with $\mathbb{E}[\tau_i]/2$, where τ_i is the total length of branches with i descendants in the gene genealogy of a sample. For an extension of independent-Poissons sampling to variants under selection, see Desai and Plotkin (2008). Our notation is different than in Seplyarskiy et al. (2021) because here we use the coalescent or diffusion time scale.

Under the standard neutral coalescent model, $\mathbb{E}[\tau_i] = 2/i$. For the general coalescent trees of Griffiths and Tavaré (1998), τ_i can be expressed in terms of the coalescent intervals, T_k, which are the lengths of time when there were $k \in (2, \ldots, n)$ lineages in the ancestry of the sample. In particular,

$$\mathbb{E}[\tau_i] = \sum_{k=2}^{n} k\mathbb{E}[T_k] \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \tag{29}$$

(Fu, 1995; Griffiths and Tavaré, 1998).

Watterson (1974a) studied three models. In Model 1, using our notation, mutations arise from a constant source at rate θ, then propagate or go extinct independently according to a critical branching process, i.e. with birth rate equal to death rate as for a neutral mutation. The number of mutations in count i has expected value $\theta\mu^i/i$, for a constant $\mu > 0$ which converges to 1 as the duration of the process increases. Watterson (1974a) proved that the numbers and counts of mutations follow the Ewens sampling formula when conditioned on their total size, which for Watterson (1974a) was equivalent to the population size. Models 2 and 3 are the Moran model and the Wright-Fisher model (Moran, 1958, 1962; Fisher, 1930b; Wright, 1931) and Watterson (1974a) proved that these have the same limit as Model 1 when the population size is large.

Model 1 is an example of a logarithmic species distribution (Fisher, 1943; Watterson, 1974b; Arratia et al., 2003; Lambert, 2011). Branching-processes have also been used to describe and infer the ages of rare alleles (Rannala and Slatkin, 1997; Slatkin and Rannala, 2000; Wiuf, 2000); for recent developments and a review, see Crespo et al. (2021). Slatkin (2000) used this approach and an extension of Griffiths and Tavaré (1998) to model the ages of rare alleles in a large sample. Champagnat and Lambert (2012, 2013) studied the convergence of population frequencies of alleles for supercritical, subcritical or critical branching processes. All of these works assume that each allele traces back to a single mutation, as under the infinite-alleles mutation model.

Our approach to modeling recurrent mutation follows that of Watterson (1974a) to Model 1. Whereas Watterson (1974a) did not specify the source of mutations, here we take it to be the production of rare variants by mutation from a common variant on the gene genealogy of a large sample. What for Watterson (1974a) was the total population size is for us the total count of a rare variant. Allele 1 is our nominal variant of interest, but for simplicity for the moment, we use n, k and θ in place of n_1, k_1 and $\theta\pi_1$. As a further notational convenience, we define

$$\bar{\tau}_i \equiv \mathbb{E}[\tau_i]$$

so that $\theta\bar{\tau}_i/2$ is the expected number of mutations with count i in this independent-Poissons sampling model.

Let $(a_1, a_2, \ldots)$ be the numbers of latent mutations of the variant of interest with counts $(1, 2, \ldots)$. We assume that $a_i \sim \text{Poisson}(\theta\bar{\tau}_i/2)$ and that a_i and a_j are independent for $i \neq j$. Their joint distribution is then

$$\begin{aligned} P(a_1, a_2, \ldots) &= \prod_{i \geq 1} \frac{(\theta\bar{\tau}_i/2)^{a_i}}{a_i!} e^{-\theta\bar{\tau}_i/2} \\ &= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \prod_{i \geq 1} \frac{(\theta\bar{\tau}_i/2)^{a_i}}{a_i!} \end{aligned} \tag{30}$$

with $a_i \geq 0$. The total sample size is what would set the upper limits of the product and the sum above, but we leave these unspecified for now, only imagining that the total sample size is much larger than the sample count of the variant of interest, so we can model the latter without restriction.

We are only concerned with a_i for $i \leq b$, where b is the largest rare-variant count. Thus, the assumption of independence in (30), which is equivalent to there being no nested mutations in the ancestry of a rare variant, will only need to be true for $\bar{\tau}_i$ with $i \in (1, \ldots, b)$. In Appendix section A.2 we prove that this holds for the trees of Griffiths and Tavaré (1998) for fixed b in the limit as the total sample size tends to infinity, and that the counts $(a_1, \ldots, a_b)$ converge to independent Poisson random variables as with expected values $(\theta\bar{\tau}_1/2, \ldots, \theta\bar{\tau}_b/2)$. A condition is that the total height of the genealogy is bounded, which is a mild assumption ruling out pathological situations such as a populations whose sizes *increase* too quickly backward in time.

The count of the variant of interest is $n = \sum_i ia_i$ and its number of latent mutations is $k = \sum_i a_i$. Following Watterson (1974a), we consider the probability generating function of n and k, which in the present case simplifies to

$$\begin{aligned} G_{n,k}(x, y) &= \sum_{(a_1, a_2, \ldots)} P(a_1, a_2, \ldots) x^n y^k \\ &= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \sum_{k=0}^{\infty} \frac{(\frac{\theta}{2})^k y^k}{k!} \Big(\sum_i x^i \bar{\tau}_i \Big)^k. \end{aligned}$$

For the details of this derivation, see (82) in the Appendix. The coefficient of x^n (and y^k) can be found using

$$\Big(\sum_i x^i \bar{\tau}_i \Big)^k = \sum_{n \geq k} x^n \sum_{(i_1, \ldots, i_{k-1})} \bar{\tau}_{i_1} \bar{\tau}_{i_2} \cdots \bar{\tau}_{i_k} \tag{31}$$

where the sum is over

$$i_m = 1, \ldots, n - (k - m) - \sum_{g=1}^{m-1} i_g$$

for $m = 1, \ldots, k - 1$, and with

$$i_k = n - \sum_{m=1}^{k-1} i_m.$$

Returning to our notation in which n_1 is the number of copies of a variant of interest, k_1 its number of latent mutations, $\theta\pi_1$ its mutation parameter, and n is the total sample size, and further using τ to show the new dependence on the vector of expected times $(\bar{\tau}_1, \ldots, \bar{\tau}_{n-1})$, we have

$$p(n_1, k_1; n \text{ large}, \tau) \approx \frac{(\frac{\theta\pi_1}{2})^{k_1} \sum_{(i_1,\ldots,i_{k_1-1})} \prod_{m=1}^{k_1} \bar{\tau}_{i_m}}{k_1!} e^{-\frac{\theta\pi_1}{2}\sum_{i=1}^{n-1}\bar{\tau}_i} \tag{32}$$

which is non-zero for $n_1 = k_1 = 0$ and $n_1 \geq k_1 \geq 1$. The sum over $(i_1, \ldots, i_{k_1-1})$ here is the same as in (31). It is equivalent to summing over partitions of the integers 1 through n_1 into k_1 subsets, where the sizes of the subsets are $(i_1, \ldots, i_{k_1})$.

It is convenient to decompose (32) as follows. The number of type-1 mutations is Poisson distributed

$$p(k_1; n \text{ large}, \tau) \approx \frac{(\frac{\theta\pi_1}{2}\sum_{i=1}^{n-1}\bar{\tau}_i)^{k_1}}{k_1!} e^{-\frac{\theta\pi_1}{2}\sum_{i=1}^{n-1}\bar{\tau}_i}, \tag{33}$$

with parameter equal to the expected number of type-1 mutations on the gene genealogy of the sample. Conditional on this, the distribution of the number of times allele 1 appears in the sample is given by

$$p(n_1|k_1; n \text{ large}, \tau) \approx \sum_{(i_1,\ldots,i_{k_1-1})} \prod_{m=1}^{k_1} \frac{\bar{\tau}_{i_m}}{\sum_{i=1}^{n-1}\bar{\tau}_i}, \tag{34}$$

which depends on the relative expected branch lengths but does not depend on θ or π_1.

Alternatively, $p(n_1; n \text{ large}, \tau)$ can be computed by summing (32) appropriately, over $k_1 \in (0, \ldots, n_1)$. Then

$$p(k_1|n_1; n \text{ large}, \tau) \approx \frac{p(n_1, k_1; n \text{ large}, \tau)}{p(n_1; n \text{ large}, \tau)} \tag{35}$$

can be used to estimate the number of independent mutations which produced the observed copies a rare allele.

The sum over $(i_1, \ldots, i_{k_1-1})$ in (34) and (32) is straightforward to evaluate but will become impractical if n_1 and k_1 become too large. In what follows, we consider $k_1 \leq 7$ mutations at a each site. Equation (33) suggests that this will be accurate up to about three expected mutations per site, because the probability of k_1 greater than 7 is just over 1% when $(\theta\pi_1/2)\sum_{i=1}^{n-1}\bar{\tau}_i = 3$. As in Figure 2, the largest value of n_1 we consider is 40. These are not the upper limits of feasibility; it takes two minutes to evaluate (34) for all $k_1 \in (0, \ldots, 7)$ and $n_1 \in (0, \ldots, 40)$ in Mathematica version 11.2 (Wolfram Research, Inc., 2017) on a mid-2015 MacBook Pro.

Considering the first three possible values of k_1 in (34),

$$p(n_1|0; n \text{ large}, \tau) \approx \begin{cases} 1 & \text{if } n_1 = 0 \\ 0 & \text{if } n_1 \geq 1 \end{cases} \tag{36}$$

$$p(n_1|1; n \text{ large}, \tau) \approx \frac{\bar{\tau}_{n_1}}{\sum_{i=1}^{n-1}\bar{\tau}_i} \tag{37}$$

$$p(n_1|2; n \text{ large}, \tau) \approx \frac{\sum_{i=1}^{n_1-1}\bar{\tau}_i\bar{\tau}_{n_1-i}}{\left(\sum_{i=1}^{n-1}\bar{\tau}_i\right)^2} \tag{38}$$

Equation (36) says simply that if there are no type-1 mutations on the gene genealogy then no copies of allele-1 will be observed. Equation (37) is the familar result for the site-frequency spectrum, that it is given by the proportion of branches in the tree that have n_1 descendants. Equation (38) extends this to two mutations and emphasizes that mutations in the ancestry of a rare allele will be non-nested when n is large.

For the constant-size model, we can put $\bar{\tau}_i = 2/i$ in (32) and obtain new expressions corresponding to (26), (27) and (28),

$$p(n_1; n \text{ large}) \propto \frac{(\theta\pi_1)_{(n_1)}}{n_1!} e^{-\theta\pi_1 \sum_{i=1}^{n-1} 1/i} \tag{39}$$

$$p(n_1; k_1, n \text{ large}) \propto \frac{S_{n_1}^{(k_1)} k_1!}{n_1!} \left(\sum_{i=1}^{n-1} \frac{1}{i} \right)^{-k_1} \tag{40}$$

$$p(n_1, k_1; n \text{ large}) \propto \frac{S_{n_1}^{(k_1)} (\theta\pi_1)^{k_1}}{n_1!} e^{-\theta\pi_1 \sum_{i=1}^{n-1} 1/i} \tag{41}$$

which may be preferable to the previous ones. The expression for $p(k_1|n_1; n \text{ large})$ obtained using $\bar{\tau}_i = 2/i$ here is the identical to (25). Figure 2 is also unchanged if (39) and (40) are used instead of (26) and (27), and normalizing in the same way.

2.1 Relation to K-alleles diffusion results

These new results may be discerned in the sampling probabilities from the diffusion model. For example, a more detailed treatment of (6) and application of Stirling's formula gives

$$\begin{aligned} p(n_1; n) &= \pi_2 \frac{(\theta\pi_1)_{(n_1)}}{n_1!} \frac{\Gamma(1+\theta)}{\Gamma(1+\theta\pi_2)} \frac{\Gamma(n+1)\Gamma(n-n_1+\theta\pi_2)}{\Gamma(n+\theta)\Gamma(n-n_1+1)} \\ &= \pi_2 \frac{(\theta\pi_1)_{n_1}}{n_1!} \frac{\Gamma(1+\theta)}{\Gamma(1+\theta\pi_2)} e^{-\theta\pi_1 \log(n)} \left[1 + O\left(\frac{1}{n}\right) \right] \end{aligned}$$

in which we write $e^{-\theta\pi_1 \log(n)}$ in place of $n^{-\theta\pi_1}$ to emphasize the connection to the gene genealogy. Using a Taylor series approximation for the Gamma function around 1 we have

$$\frac{\Gamma(1+\theta)}{\Gamma(1+\theta\pi_2)} = 1 - \gamma\theta\pi_1 + O(\theta^2) \approx e^{-\gamma\theta\pi_1}$$

where $\gamma = 0.577\ldots$ is Euler's constant. As $\sum_{i=1}^{n} 1/i \approx \log(n) + \gamma$ when n is large, we find that (6) will be very close to π_2 times the right-hand side of (39) for a given n_1 when n is large and θ is small. For reference, note that even the fastest-rate sites in the human genome have θ only equal to about 0.02 (see Section 3.2 below).

In deriving (32), we assumed that both the number of mutations and their total count (k_1 are n_1 here) are unbounded. We did not take a formal limit as $n \to \infty$, and instead have $(\theta\pi_1/2) \sum_{i=1}^{n-1} \bar{\tau}_i$ for the rate of occurence of type-1 mutations. This is intuitive from the standpoint of coalescent theory, since $\sum_{i=1}^{n-1} \bar{\tau}_i$ is the average total length of the gene genealogy, but allowing that k_1 are n_1 could potentially be larger than n makes little sense. These new results (32) through (41) are for rare alleles only, that is for a given n_1 when n is very large.

To relate these new results to the model in Section 1.1, we can follow the logic in Section 1.1.1 and approximate the full sampling probability (6) when θ is small as the sum of two copies of the independent-Poissons sampling process. Thus, similar to (10), for the two-allele case here we can

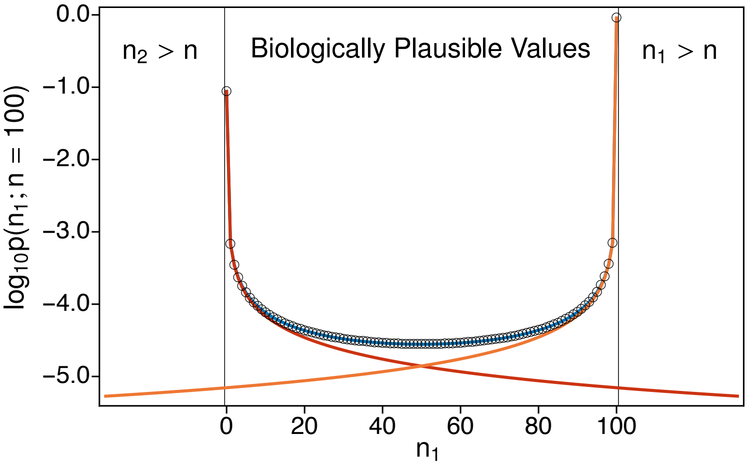


Figure 3: Schematic relating the independent-Poissons model to the classic two-allele diffusion result (6). Parameters are as in Figure 1, specifcally $\theta\pi_2 = 1/1300$, $\pi_1 = 10\pi_2$ and $n = 100$. The blue line shows the sum (42). The red line shows the first term of that sum. The orange line shows the second term. The plot extends beyond the biologically plausible values of n_1 because the independent-Poissons model allows $n_1 > n$ and $n_2 > n$. Open circles show (6) which closely tracks the blue line across the range of biologically plausible values, and the red and orange lines for small n_1 and small n_2 respectively. Biologically implausible are rare in this case.

approximate the sampling probability $p(n_1; n)$ in (6) using

$$\pi_2 \frac{(\theta\pi_1)_{(n_1)}}{n_1!} e^{-\theta\pi_1 \sum_{i=1}^{n-1} 1/i} + \pi_1 \frac{(\theta\pi_2)_{(n_2)}}{n_2!} e^{-\theta\pi_2 \sum_{i=1}^{n-1} 1/i} \tag{42}$$

in which we have used result (39) twice, once for type-1 mutations in a type-2 ancestral background and once for type-2 mutations in an type-1 ancestral background, where it is understood that $\pi_2 = 1 - \pi_1$ and $n_2 = n - n_1$. This can be seen as a sample-based version of the boundary mutation model (Vogl and Clemente, 2012; Schrempf and Hobolth, 2017; Vogl et al., 2020) but one which allows for multiple segregating mutations.

Figure 3 plots the two terms in (42) individually, together with their sum (42) and $p(n_1; n)$ from (6). The parameters as the same as the small-θ case in Figure 1, specifically a sample of size $n = 100$ with θ chosen so that the mutation rate for allele 2 ($\theta\pi_2$) is equal to the human average (1/1300) and $\pi_1 = 10\pi_2$. In contrast to Figure 1, the range of n_1 (similarly n_2) in Figure 3 extends beyond what is biologically plausible. For these parameters, the expected number of type-1 mutations on the gene genealogy is about 0.04 and the expected number of type-2 mutations is about 0.004. Only about 1/50 polymorphic sites where the rare variant is allele 1 are expected to have experienced more than one mutation, and the corresponding value for sites where the rare variant is allele 2 is about 1/500. The different probabilites at the boundaries 0 and 100 result both from the ten-fold greater rate of type-1 versus type-2 mutation on the gene genealogy and the weights π_2 and π_1 in (42) which similarly capture the ten-fold difference in the the chance of the ancestral allele being of type 2 or of type 1, respectively, at stationarity in the Wright-Fisher diffusion model.

3 Theoretical example and data application

Here we illustrate the theoretical and empirical use of (33) and (34). First we describe the consequences of recurrent mutation in an exponentially growing population compared to those in a population of constant size. Second we explore an entirely empirical application to human SNP data, which suggests that disparate site-frequency spectra may be explained by differences in mutation rate (and thus recurrent mutation).

Note that if estimates of the expected fraction of the gene genealogy comprised of branches with i descendants, that is

$$\frac{\bar{\tau}_i}{\sum_{i=1}^{n-1}\bar{\tau}_i} = \frac{\mathbb{E}[\tau_i]}{\sum_{i=1}^{n-1}\mathbb{E}[\tau_i]}, \tag{43}$$

are available, then $p(n_1|k_1; n \text{ large}, \tau)$ can be computed using (34). In addition, for any estimated or supposed values of the expected number of mutations on the gene genealogy,

$$\frac{\theta\pi_1}{2}\sum_{i=1}^{n-1}\bar{\tau}_i = \frac{\theta\pi_1}{2}\sum_{i=1}^{n-1}\mathbb{E}[\tau_i], \tag{44}$$

the joint distribution of the number of latent mutations, k_1, and their total count, n_1, is the product of (33) and (34).

3.1 An exponentially growing population

Consider the simple model of pure exponential growth which has been the subject of a number of studies (Slatkin and Hudson, 1991; Griffiths and Tavaré, 1998; Polanski and Kimmel, 2003; Chen and Chen, 2013; Polanski et al., 2017): a population which has reached its current (haploid) size N_0 by exponential growth at rate r per generation. On the coalescent time scale of N_0 generations, looking backward in time and setting $\alpha = N_0 r$,

$$N(t) = N_0 e^{-\alpha t} \tag{45}$$

gives the population size at time t in the past. This model is unrealistic because the past population size approaches zero, but it can be taken as a rough approximation for recent dramatic growth. For instance, a population of current size $N_0 = 5 \times 10^7$ with a generation time is 30 years and $r = 0.0064$, would have $\alpha = 3.2 \times 10^5$. About 40,000 years ago, it would have had size 10^5, and using equation (7) in Slatkin and Hudson (1991) the pairwise coalescence time would be about 57,000 years.

The expectation $\mathbb{E}[\tau_i]$ can be computed from (29) if the expected coalescent intervals $\mathbb{E}[T_k]$ are known. We use the large-n results of Chen and Chen (2013) for $\mathbb{E}[T_k]$ (our notation) to obtain a simple approximation for $\mathbb{E}[\tau_i]$. With the time scale and notation here, equation (11) in Chen and Chen (2013) gives

$$\frac{1}{\alpha}\log\left(2\alpha\left(\frac{1}{k}-\frac{1}{n}\right)+1\right) \tag{46}$$

as a large-n approximation for the cumulative expected time for the number of ancestral lineages of the sample to decrease from n to k. Writing (46) as a continuous function of $x = k/n$,

$$f(x) = \frac{1}{\alpha}\log\left(\frac{2\alpha}{n}\frac{1-x}{x}+1\right), \tag{47}$$

we approximate the expected coalescent interval as

$$\begin{aligned}\mathbb{E}[T_k] &= f(x-dx) - f(x) \approx -f'(x)dx \\ &= \frac{2}{x\left(2\alpha(1-x)+xn\right)}.\end{aligned} \tag{48}$$

Note that while (48) is a large-n approximation, it allows that α might be of the same order of magnitude as n. Applying the same approximation to the combinatorial coefficient in (29) gives

$$\frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \approx \frac{x}{1-x}(1-x)^i. \tag{49}$$

Finally, we approximate the sum in (46) with the integral

$$\begin{aligned}\mathbb{E}[\tau_i] &\approx \int_0^1 xn \frac{2}{x\left(2\alpha(1-x)+xn\right)} \frac{x}{1-x}(1-x)^i dx, \\ &= 2\int_0^1 \left[1-(1-2\alpha/n)y\right]^{-1}(1-y)y^{i-1}dy \end{aligned} \tag{50}$$

in which we changed variables ($y = 1 - x$) to make a connection to the hypergeometric function. Thus we obtain

$$\mathbb{E}[\tau_i] \approx \frac{2}{i(i+1)}\ {}_2F_1(1, i; i+2; 1-2\alpha/n). \tag{51}$$

Equation (51) can be evaluated efficiently and the properties of the hypergeometric function are well know.

The suggested dependence of (51) on $1/i^2$, rather than the usual site-frequency prediction of $1/i$, is consistent with the excess of low-frequency variants expected under population growth. As Slatkin and Hudson (1991) and others have observed, gene genealogies under very fast exponential growth are close to star trees. In this extreme, all variants will be singletons. From (51), when α/n is large, we have

$$\mathbb{E}[\tau_i] \approx \begin{cases} \frac{\log(2\alpha/n)-1}{\alpha/n} & \text{if } i = 1 \\ \frac{2}{i(i-1)\alpha/n} & \text{if } i \geq 2 \end{cases} \tag{52}$$

if terms of order $(\alpha/n)^{-2}$ or smaller are ignored. Thus singletons will dominate as expected when α/n is very large.

These results for exponentially growing populations, derived here using a coalescent approach, are identical in form to some results for "Luria-Delbrück distributions", especially in application to cancer, derived using forward-time birth-death or branching processes (Luria and Delbrück, 1943; Lea and Coulson, 1949; Kessler and Levine, 2013; Ohtsuki and Innan, 2017; Gunnarsson et al., 2021; Cheek and Antal, 2018; Poon et al., 2021; Durrett, 2013, 2015). In particular, (50) has the same form as the approximation in equation (4) of Ohtsuki and Innan (2017) and as equation (33) in Gunnarsson et al. (2021). Equation (52) has the same form as the expression in Theorem 2 in Durrett (2013) if only the leading-order term is kept in (52) in the case $i = 1$.

Figure 4 shows the same quanties as Figure 2 but for the pure exponential growth model with $n = 10^5$ and $\alpha/n = 3$. The value $\alpha/n = 3$ was chosen to roughly reproduce the ratio of singletons to doubletons observed for low-rate sites in the gnomAD data in Section 3.2. Figure 4A is directly comparable to Figure 2A, the only difference being whether $\mathbb{E}[\tau_i] = 2/i$ or comes from (51). As Figure 4A shows, recent rapid growth produces a single-mutation ($k_1 = 1$, blue line) site-frequncy spectrum with an excess of rare variants and a deficit of common variants. So, compared to the constant-size case in Figure 2A, there is a diminished tendency to observe high-frequency variants when the number of latent mutations is larger, and a stronger tendency for the site-frequency count (n_1) to be equal to or close to the number of latent mutations.

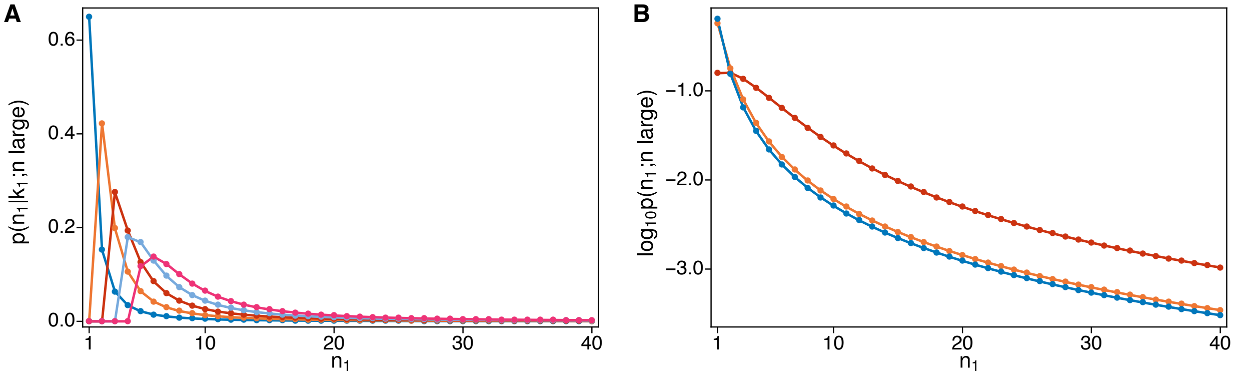


Figure 4: Plots of the same quantities shown in Figure 2 but for a sample of size $n = 10^5$ under pure exponential growth with $\alpha/n = 3$. Panel A: probability of observing n_1 copies of allele 1 in the sample given $k_1 = 1, 2, 3, 4, 5$ latent mutations. Panel B: $\log_{10}$-probability of observing n_1 copies of allele 1 in the sample for three different mutation rates, corresponding to the values of $\theta\pi_1$: 0.002, 0.02 and 0.2 in Figure 2, but here expressed in terms of expected numbers of mutations on the gene genealogy (44): 0.024, 0.24 and 2.4. Probabilities in both panels are normalized to sum to one for $n_1 \in (1, 2, \ldots, 40)$.

To make Figure 4B comparable to Figure 2B, we used (44) with $n = 10^5$ and $\mathbb{E}[\tau_i] = 2/i$ to compute the corresponding expected numbers of mutations on the gene genealogy for the three values of $\theta\pi_1$ in Figure 2 $(0.002, 0.02, 0.2)$. The resulting expected numbers of mutations were 0.024, 0.24 and 2.4, the last being about equal to the value for the highest-rate sites in the gnomAD data in Section 3.2. We then computed $p(n_1; n \text{ large}, \tau)$ by averaging (34) over the distribution (33). Similar to Figure 2B, the two smaller values of the mutation rate give nearly indistinguishable results for the total count n_1. But there is a dramatic difference for the largest mutation rate. In Figure 2B the prediction is distinctly L-shaped and thus similar to that for the lowest mutation rate, which again is 100-fold lower. In contrast, in Figure 4B singletons have a much lower chance of being observed. In fact, doubletons are slightly more likely than singletons. This relative excess of doubletons is due to the fact when there are two latent mutations these are highly likely to produce two copies of the variant under growth (Figure 4A) than under constant size (Figure 2A).

It is also of interest to know how the number of latent mutations in the ancestry of a rare variant depends on its count. Figure 5 depicts this for a series of increasing counts n_1, from 1 to 16. Figure 5A shows the results for constant size, Figure 5B the corresponding results for pure exponential growth. The expected number of mutations on the gene genealogy is 2.4 in both cases. Regardless of the demography, if only one copy of the variant is observed, it must be due to one mutation. Otherwise, the results differ greatly for constant size versus growth. Under constant size, a variant observed multiple times in the sample can easily be due to a single mutation. Under growth, higher variant counts are more likely due to multiple mutations.

3.2 Application to human SNP data

We also used (33) and (34) to account for latent mutations in the ancestry of low-count variants in a subset of the gnomAD data (Karczewski et al., 2020). We took the approach described in the Supplementary Materials of Seplyarskiy et al. (2021), specifically obtaining estimates of relative branch lengths (43) from low-rate sites then using our new analytical result (34) to average over mutation counts. Rather than categorizing variants by trinucleotide context as in Seplyarskiy et al. (2021), we analyzed data from gnomAD version v2.1.1, presorted into 109 bins based on estimates of

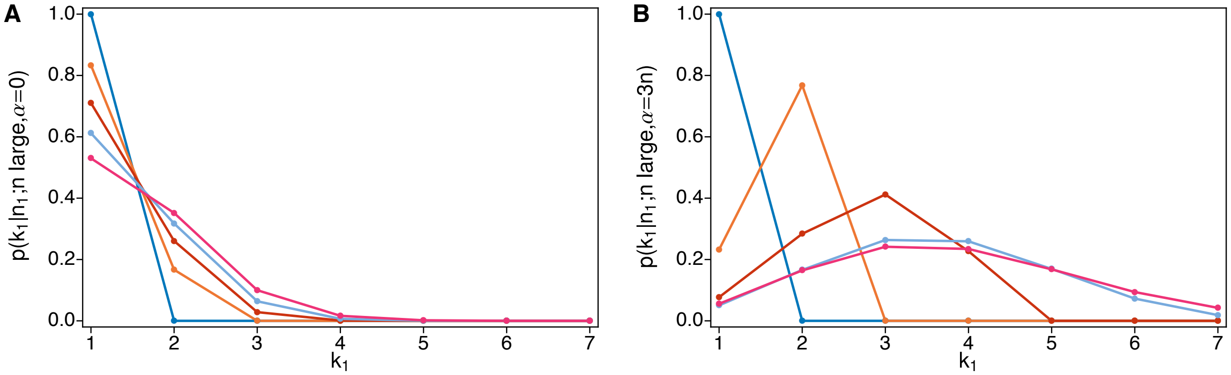


Figure 5: Probabilities of $k_1 = 1, 2, 3, 4, 5, 6, 7$ latent mutations for increasing values of n_1 — 1, 2, 4, 8, and 16, with blue for 1, orange for 2, and so on — when 2.4 mutations are expected on the gene genealogy of a sample of size $n = 10^5$ (or equivalently $\theta\pi_1 = 0.2$ in the constant-size case). Panel A plots (25) with $\theta\pi_1 = 0.2$. Panel B shows the same probability computed using (33) and (34) under exponential growth with $\alpha/n = 3$.

mutation rate by the method of Seplyarskiy et al. (2022, in prep.) which incorporates information from the six flanking bases on either side of a SNP, strand asymmetry, expression level, methylation promoter status. We did not use this information but our analysis assumes that variants within a bin have the same mutation rate.

The data consist of variant counts for synonymous mutations in the exomes of about 57K non-Finnish Europeans. Thus $n \sim$ 114K although this varied by about 2% among sites because we required that sites were successfully genotyped in a minimum of 112K chromosomes. Importantly for our application, the data include monomorphic sites, i.e. sites with variant count equal to zero. gnomAD only provides n for polymorphic sites, so we imputed n for monomorphic sites using the nearest value at a polymorphic site within 100bp on either side of the focal site. After filtering for sequencing quality and coverage as well as removing mutation rate bins with fewer than 100 observed mutations, there are a total of $12,338,176$ sites in 97 bins and $834,486$ of these are polymorphic.

Figure 6 shows the total numbers of sites (blue circles) and the number of monomorphic sites (orange circles) in each bin. The great majority of sites are in bins 1 through roughly 25. These have low mutation rates, as indicated by the close overlap of blue and orange circles, or nearly equal numbers of total sites and monomorphic sites. The widening gap between the blue and orange circles reflects the fact that higher-number bins have larger mutation rates. Estimates of the expected numbers of mutations per site for each bin range from 0.0097 for bin 1 to 2.23 for bin 97, with a mean of 0.083 (see below).

Each bin contains a mixture of different sequence contexts and different mutations. Again, we assume that within a bin these all have the same rate. We use $\theta\pi_1$ to denote this rate. Let S_i be the number of sites in a given bin where i copies of a variant are observed in the sample. If a bin contains L total sites, then with reference to the notation in (2) we may write

$$\mathbb{E}[S_i] = L\mathbb{P}\left[\mathcal{N}_1 = i; n\right], \quad i \in (0, \ldots, n-1). \tag{53}$$

Thus we use i in place of n_1 to avoid the additional subscript when we apply the results of the previous sections.

We compare observed and expected site-frequency counts for each bin based on an entirely empirical fit of our general results (33) and (34). This involves three steps. First, we use (33)

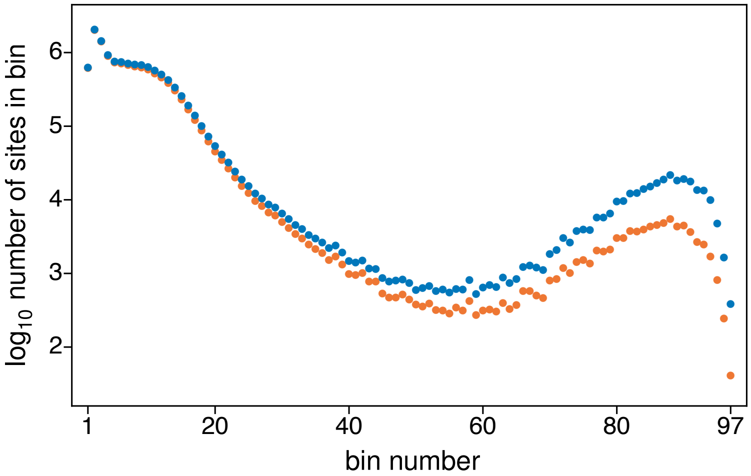


Figure 6: Blue circles show the total numbers of sites in the gnomAD data for each of the 97 bins. Orange circles show show the total numbers of monomorphic sites, i.e. sites with variant count equal to zero, in each bin. Lower mutation-rate bins are on the left, higher mutation-rate bins are on the right. The estimated mutation rates for bins 1 and 97 and about 9 times smaller and 27 times larger than the average (see text).

and the proportion of monomorphic sites, S_0/L, to estimate the total mutation rate in each bin, $\theta\pi_1 \sum_i \bar{\tau}_i$, i.e as $-\log(S_0/L)$. Next, based on (37), we use data for low-rate bins to estimate $\bar{\tau}_i / \sum_i \bar{\tau}_i$ as $S_i/(L - S_0)$ for $i \in (1, \ldots, 40)$. Finally, we compute the expectations $\mathbb{E}[S_i]$, $i \in (1, \ldots, 40)$, for each bin using the estimated total mutation rate and the total number of sites for that bin. We do this using (33) and (34), and assuming that the $\bar{\tau}_i / \sum_i \bar{\tau}_i$ estimated from low-rate bins holds for all bins.

We used the combined variant counts for the first five bins to estimate the relative branch lenghts $\bar{\tau}_i / \sum_i \bar{\tau}_i$. Our estimates of the total mutation rate for these bins range from 0.0097 to 0.037 with an average of 0.021. This is somewhat less than the smallest mutation rate in Figure 4 (also Figure 2) from which we can infer that these sites are unlikely to be affected by multiple mutations. Again, over all 97 bins, our estimates range from 0.0097 to 2.23, or about 230-fold from lowest to highest. The average across all bins is 0.083. Assuming that the latter corresponds to the genome average mutation rate per site, for which the usual estimate of θ from pairwise differences is about $1/1300 \sim 0.0077$, we can infer that the expected number of mutations between a pair of (haploid) genomes is about 9×10^{-5} for the slowest sites and about 0.02 for the fastest sites.

Figure 7 compares the observed and expected variant counts, S_i for $i \in (1, \ldots, 40)$, for bins 9, 50 and 92, chosen to represent a low-rate bin, a middle-rate bin and a high-rate bin. Figure 9 in the Appendix gives the plots for all 97 bins. Red circles show the observed counts. Blue lines trace the expected values. In making these plots, we grouped variant counts for which $\mathbb{E}[S_i] < 1$. For bin 50 for example, this was true of variant counts $i \in [12, 40]$ as noted in Figure 7B and again in the 50th panel of Figure 9. The values of "mutrate" displayed in these plots are the estimates of the expected number of mutations per site on the gene genealogy, i.e. the total mutation rate $\theta\pi_1 \sum_i \bar{\tau}_i$, for each bin based on its proportion monomorphic sites.

The broad pattern from these plots is clear. For small total mutation rates (e.g. Figure 7A) the site-frequency spectrum is heavily weighted toward rare variants. For large total mutation rates (e.g. Figure 7C), i.e. when multiple latent mutations are likely, the site-frequency spectrum is shifted toward higher-frequency variants. As shown in Figure 6, the data contain fewer sites with intermediate mutation rates. In this case (e.g. Figure 7B), the site-frequency spectrum does show

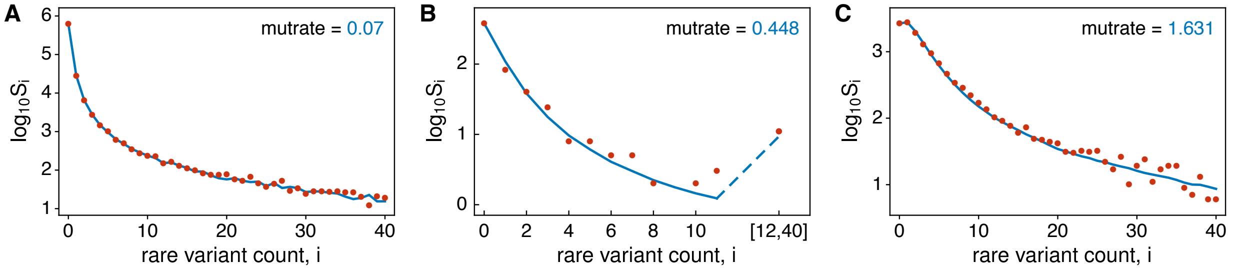


Figure 7: Examples of model fit for a (A) low-rate bin, (B) middle-rate bin, and (C) high-rate bin. Red dots are the data. Blue lines show expectations with "mutrate" $\theta\pi_1 \sum_i \bar{\tau}_i$ estimated using the new method. These are bins 9, 50 and 92 (cf. Figure 6).

the expected intermediate pattern, but subject to considerable sampling error. Across the range of mutation rates, the empirical model, which uses low-rate sites to estimate relative branch lengths $\bar{\tau}_i / \sum_i \bar{\tau}_i$ and assumes these hold for all sites, fits the data well.

As can be seen in Figure 7A and the first 20 or so panels of Figure 9, the empirical estimates of $\bar{\tau}_i / \sum_i \bar{\tau}_i$ include fluctuations due to sampling error for higher-count variants. The combined data for the first five bins have S_i ranging from 71 to 38 for $i \in [30, 40]$. The presence of these fluctuations helps illustrate a subtler phenomenon, namely the smoothing which occurs at larger mutation rates (e.g. Figure 7C). For reference, the combined data for the first five bins have S_i in the thousands for the low-count variants. From these, the estimated chance that a latent mutation is a singleton is about 64%, followed by 13% for doubletons and 6% for tripletons. By comparison, the chance is less than 0.1% for each variant with count $i \in [25, 40]$. The predictions $\mathbb{E}[S_i]$ are smoothed for higher-count variants at larger mutation rates because they are mixtures. For example, two latent mutations will come in counts 1 and $i-1$, 2 and $i-2$, or 3 and $i-3$ with approximate relative proportions 64:13:6.

4 Discussion

In this work, we modeled the mutational ancestry of a rare variant in a large sample. Under the standard neutral model of population genetics with K-allele parent-independent mutation, we found that co-segregating rare variants may be treated independently and that the Ewens sampling formula gives the probabilistic structure of latent mutations in their ancestries. We obtained more general results, which hold under changing population size, by modeling latent mutations as independent Poisson random variates.

Our aim has been to describe how the site-frequency spectra of rare variants in large samples are affected recurrent mutation. The key parameters for a variant in count i turn out to be its expected total rate of mutation on the gene genealogy of the sample (here denoted $\theta\pi_1 \sum_i \bar{\tau}_i$) and the expected relative lengths of branches in the gene genealogy which have i descendants ($\bar{\tau}_i / \sum_i \bar{\tau}_i$). Under the standard neutral model $\bar{\tau}_i = 2/i$.

We obtained new results for $\bar{\tau}_i$ under exponential population growth and used these to illustrate how recurrent mutation affects the site-frequency spectrum differently than under constant size. Lastly, we showed that our general results provide a good fit to synonymous variation among a large number of (non-Finish European) individuals in the human Genome Aggregation Database (Karczewski et al., 2020), suggesting that, whatever the causes of deviations from $\bar{\tau}_i = 2/i$ might be for this sample, differences in mutation rate can explain differences in site-frequency spectra among different kinds of sites.

Our application was empirical. We did not fit a demographic model, but following Seplyarskiy et al. (2021) used low-mutation-rate sites to estimate relative branch lengths and assumed these hold for all sites. Site-frequency spectra are a rich source of information about population-genetic phenomena but are of somewhat limited use in disentangling their effects (Myers et al., 2008; Bhaskar and Song, 2014; Terhorst and Song, 2015; Lapierre et al., 2017; Rosen et al., 2018). When low-mutation-rate sites are plentiful enough to provide stable estimates of relative branch lengths, this empirical method offers a way to control for myraid factors and isolate the effects of variation in mutation rate.

We began with a K-allele model with parent-independent mutation, and used its sampling probabilities in our computations for constant-size populations. We conjecture that our findings will hold for general mutation models because conditioning on a rare variant in a large sample means that the ancestral allele will be the common allele with very high probability. Then the relevant mutation rate in any model will be the rate of the production of the rare allele from the common allele.

We have described our general results as being for populations which may have changed in size. This is appropriate for the general coalescent model of Griffiths and Tavaré (1998) which we used to portray our results and assumed in the proofs in the Appendix. Strictly speaking, though, the general coalescent does not require a generative model for the times between coalescent events, T_k for $k \in (2, \ldots, n)$. Selection might be part of the reason they differ from the predictions of the standard neutral coalescent. This may be true, for example, for the synonymous exome data from gnomAD we analyzed.

In fact, the derivation of (33) and (34), with associated results from (30) to (38), does not even require interpretation in terms of coalescence times. These equations hold just as well if we replace $\theta\pi_1\bar{\tau}_i/2$ with an arbitrary rate parameter λ_i for the production of mutants in count i, potentially of an allele which is under selection. The case of a fixed tree, with fixed τ_i not from a generative model, considered in the Appendix is an example. The modified Poisson Random Field model of Desai and Plotkin (2008) is another, in which λ_i was the rate under additive selection (and the independent-Poissons assumption was applied to all counts in the sample). We have shown in detail how our results follow from the standard neutral coalescent or diffusion model and its extension the general coalescent model. As with our conjecture about general mutation models, we expect these results can be applied to latent mutations of alleles under various kinds of selection and a range of demographies (Lange and Fan, 1997; Dorman et al., 2004; Lambert, 2011; Kaj and Mugal, 2016; Torres et al., 2020; Müller et al., 2022) because they are for rare variants in large samples.

Acknowledgments

We thank Vladimir Seplyarskiy for helpful discussions and assistance with the mutation-rate binned SFS data from gnomAD. This work was supported in part by National Science Foundation grants DMS-1855417 and DMS-2152103, and Office of Naval Research grant N00014-20-1-2411 to Wai-Tong (Louis) Fan; and by National Institutes of Health grants R35-GM127131, R01-MH101244, U01-HG012009 and R01-HG010372 to Shamil Sunyaev.

A Appendix

A.1 Time-dependent conditional ancestral process

Here we study the conditional ancestral process in detail and provide the justification for (22) and (23).

Let $\mathcal{N}_1(t)$ and $\mathcal{N}_2(t)$ be the numbers of rare alleles and common alleles respectively at time t. From (19), (20) and (21), the stochastic process $\{(\mathcal{N}_1(t), \mathcal{N}_2(t))\}_{t\in\mathbb{R}_+}$ is a continuous-time Markov chain on $\mathbb{Z}_+^2$ with total rate of events $\lambda(n_1, n_2) = n_2^2/2$ and one-step transitions

$$(n_1, n_2) \to \begin{cases} (n_1-1, n_2+1) & \text{w/prob. } \frac{\theta\pi_1}{\theta\pi_1+n_1-1}\frac{n_1}{n_2} \\ (n_1-1, n_2) & \text{w/prob. } \frac{n_1-1}{\theta\pi_1+n_1-1}\frac{n_1}{n_2} \\ (n_1, n_2-1) & \text{w/prob. } 1-\frac{n_1}{n_2} \end{cases} \tag{54}$$

Let $\mathbb{P}_{\mathbf{n}}$ be the probability measure for this process starting at $\mathbf{n} = (n_1, n_2)$, and define the random times

$$\mathcal{T}_i := \inf\{t \ge 0 : \mathcal{N}_1(t) = n_1 - i\} \tag{55}$$

to be the times at which the first coordinate of the process decreases to $n_1 - i$ for $1 \le i \le n_1$, with $\mathcal{T}_0 = 0$. We have $0 = \mathcal{T}_0 < \mathcal{T}_1 < \mathcal{T}_2 < \cdots < \mathcal{T}_{n_1}$ almost surely under $\mathbb{P}_{\mathbf{n}}$, and the process $(\mathcal{N}_1, \mathcal{N}_2)$ visits the following points in order $(n_1, n_2) \to (n_1 - 1, \mathcal{N}_2(\mathcal{T}_1)) \to \cdots\cdots \to (0, \mathcal{N}_2(\mathcal{T}_{n_1}))$.

In Theorem 1 we describe the joint distribution of the hitting times $(\mathcal{T}_i)_{i=1}^{n_1}$ and the locations $(\mathcal{N}_2(\mathcal{T}_i))_{i=1}^{n_1}$ as $n_2 \to \infty$.

Theorem 1. *As $n_2 \to \infty$, the random vector*

$$\left(n_2(\mathcal{T}_i - \mathcal{T}_{i-1}),\ \frac{\mathcal{N}_2(\mathcal{T}_i)}{n_2}\right)_{i=1}^{n_1} \tag{56}$$

in $\mathbb{R}_+^{2n_1}$ converges in distribution under $\mathbb{P}_{\mathbf{n}}$ to the random vector

$$\left(\frac{Z_i}{(1-Y_0)(1-Y_1)\cdots(1-Y_{i-1})},\ (1-Y_1)\cdots(1-Y_i)\right)_{i=1}^{n_1},$$

where $Y_0 = 0$, and $\{Y_i,\ Z_i\}_{i=1}^{n_1}$ are independent random variables with probability density functions

$$\begin{aligned} f_{Y_i}(y) =& (n_1-i+1)(1-y)^{n_1-i} \qquad \text{for} \quad y \in (0,1) \qquad \text{and} \\ f_{Z_i}(z) =& (n_1-i+1)\,\frac{2^{n_1-i+1}}{(z+2)^{n_1-i+2}} \quad \text{for} \quad z \in (0,\infty). \end{aligned}$$

Remark 1 (Mean of $\mathcal{T}_{n_1}$)**.** Note that

$$\mathbb{E}[Z_i] = \begin{cases} \frac{2}{(n_1-i)} & \text{if } 1 \le i \le n_1 - 1 \\ \infty & \text{if } i = n_1 \end{cases}.$$

Hence for $n_1 \ge 2$, Theorem 1 implies that $\mathbb{E}_{\mathbf{n}}[\mathcal{T}_1]$ is of order $1/n_2$ and gives the second part of (23) in the main text. In contrast, when $n_1 = 1$, $\mathbb{E}[Z_1] = \infty$ and $\mathbb{E}_{\mathbf{n}}[\mathcal{T}_{n_1}]$ is no longer of order $1/n_2$.

Indeed, when $n_1 = 1$,$\mathbb{P}_{\mathbf{n}}(\sharp = k) = \frac{1}{n_2}$ for $k \in \{0, 1, \cdots n_2 - 1\}$ by (59). Hence by (57) and Fubinni's theorem,

$$\begin{aligned}
\mathbb{E}_{\mathbf{n}}[\mathcal{T}_1] &= \sum_{k=0}^{n_2-1} \sum_{i=0}^{k} \mathbb{E}_{\mathbf{n}}[\xi_i]\, \mathbb{P}_{\mathbf{n}}(\sharp = k) \\
&= \sum_{k=0}^{n_2-1} \sum_{i=0}^{k} \frac{2}{(n_2 - i)^2} \frac{1}{n_2} \\
&= \frac{2}{n_2} \sum_{i=0}^{n_2-1} \frac{1}{(n_2 - i)} \\
&\approx \frac{2}{n_2} \log n_2 \qquad \text{as } n_2 \to \infty.
\end{aligned}$$

These give (22) in the main text.

Remark 2 (Mean of $\mathcal{N}_2(\mathcal{T}_1)$)**.** By (58) and Theorem 1,

$$\lim_{n_2\to\infty} \mathbb{E}\left[\frac{\mathcal{N}_2(\mathcal{T}_i)}{n_2}\right] = \frac{n_1}{n_1+1} \frac{n_1-1}{n_1} \cdots \frac{n_1-i+1}{n_1-i+2} = \frac{n_1-i+1}{n_1+1}$$

for $1 \le i \le n_1$. This gives (23) in the main text.

A.1.1 Proof of Theorem 1

To explain the key idea we first establish weak convergence of $\left(n_2\, \mathcal{T}_1,\, \frac{\mathcal{N}_2(\mathcal{T}_1)}{n_2}\right)$, i.e. of the marginal distribution for $i = 1$ in (56). By definition, $\mathcal{T}_1$ is given by

$$\mathcal{T}_1 = \sum_{i=0}^{\sharp} \xi_i, \tag{57}$$

where $\sharp$ is the number of downward jumps in second coordinate of the process starting at (n_1, n_2) up to the first decrease in the first coordinate. The variables $\{\xi_i\}_{i=0}^{\sharp-1}$ are the times between these downward jumps, with $\xi_\sharp$ being the time to the final jump starting at $(n_1, n_2 - \sharp)$. This last jump is the one which decreases the first coordinate. Observe that $\mathcal{N}_2(\mathcal{T}_1)$ is either $n_2 - \sharp$ or $n_2 - \sharp + 1$. Given $\sharp$, $\mathcal{N}_2(\mathcal{T}_1)$ is equal to

$$\begin{cases} n_2 - \sharp + 1, & \text{w/conditional prob. } \frac{\theta\pi_1}{\theta\pi_1+n_1-1} \frac{n_1}{n_2-\sharp} \\ n_2 - \sharp, & \text{w/conditional prob. } \frac{n_1-1}{\theta\pi_1+n_1-1} \frac{n_1}{n_2-\sharp} \end{cases} \tag{58}$$

which correspond to a non-empty mutation event and a coalescent event of type 1 respectively. These follow from (54).

The probability mass function of $\sharp$ is given by $\mathbb{P}_{\mathbf{n}}(\sharp = 0) = \frac{n_1}{n_2}$ and, for $k \in \{1, 2, \cdots, n_2 - n_1\}$,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{n}}(\sharp = k) \\
&= \left(1 - \frac{n_1}{n_2}\right)\left(1 - \frac{n_1}{n_2 - 1}\right) \cdots \left(1 - \frac{n_1}{n_2 - k + 1}\right) \frac{n_1}{n_2 - k} \\
&= \frac{n_1}{n_2} \prod_{j=1}^{n_1-1} \frac{n_2 - k - j}{n_2 - j} \qquad (59) \\
&\approx \frac{n_1}{n_2} (1 - x)^{n_1-1} \qquad (60)
\end{aligned}$$

as $n_2 \to \infty$ and $\frac{k}{n_2} \to x \in (0,1)$. Hence $\mathbb{P}_{\mathbf{n}}(\sharp > n_2 - n_1) = 0$ and, for $k \in \{0, 1, 2, \cdots, n_2 - n_1 - 1\}$,

$$\begin{aligned}
\mathbb{P}_{\mathbf{n}}(\sharp > k) &= \left(1 - \frac{n_1}{n_2}\right)\left(1 - \frac{n_1}{n_2 - 1}\right) \cdots \left(1 - \frac{n_1}{n_2 - k}\right) \\
&\approx \prod_{j=1}^{n_1} \left(1 - \frac{k+j}{n_2}\right) \\
&\approx (1-x)^{n_1}
\end{aligned} \tag{61}$$

as $n_2 \to \infty$ and $\frac{k}{n_2} \to x \in (0,1)$.

Lemma 1. *As $n_2 \to \infty$, we have convergence in distribution*

$$\left(n_2 \sum_{i=0}^{\sharp} \xi_i,\ \frac{\sharp}{n_2}\right) \xrightarrow{\mathcal{L}} (Z_1,\ Y_1)\,.$$

with Z_1 and Y_1 as defined in Theorem 1.

Proof of Lemma 1. It suffices to show that

$$\lim_{n_2 \to \infty} \mathbb{E}_{\mathbf{n}}\left[e^{\eta \frac{\sharp}{n_2} + \zeta\, n_2 \mathcal{T}_1}\right] = n_1 \int_0^1 (1-x)^{n_1 - 1}\, e^{\{\eta\, x + \frac{2\zeta x}{1-x}\}}\, dx \tag{62}$$

for $\eta \in \mathbb{R}$ and $\zeta \in (-\infty, 0]$. Since $\xi_i \sim \mathrm{Exp}(\lambda(n_1, n_2 - i))$,

$$\mathbb{E}_{\mathbf{n}}\left[e^{\zeta\, \xi_i}\right] = \frac{\lambda(n_1, n_2 - i)}{\lambda(n_1, n_2 - i) - \zeta} = \frac{(n_2 - i)^2}{(n_2 - i)^2 - 2\zeta}. \tag{63}$$

By (57), (59) and (63),

$$\begin{aligned}
\mathbb{E}_{\mathbf{n}}\left[e^{\eta \frac{\sharp}{n_2} + \zeta\, n_2 \mathcal{T}_1}\right] &= \sum_{k=0}^{n_2 - n_1} \mathbb{P}_{\mathbf{n}}(\sharp = k)\, e^{\eta \frac{k}{n_2}}\, \mathbb{E}_{\mathbf{n}}\left[e^{\zeta\, n_2 \sum_{i=0}^{k} \xi_i}\right] \\
= \sum_{k=0}^{n_2 - n_1} \mathbb{P}_{\mathbf{n}}(\sharp = k)\, e^{\eta \frac{k}{n_2}} &\prod_{i=0}^{k} \mathbb{E}_{\mathbf{n}}\left[e^{\zeta\, n_2\, \xi_i}\right] \\
= \frac{n_1}{n_2} \sum_{k=0}^{n_2 - n_1} e^{\eta \frac{k}{n_2}} &\prod_{j=1}^{n_1 - 1} \frac{n_2 - k - j}{n_2 - j}\, p_{n_2}(\zeta),
\end{aligned} \tag{64}$$

where

$$\begin{aligned}
p_{n_2}(\zeta) &:= \prod_{i=0}^{k} \frac{\lambda(n_1, n_2 - i)}{\lambda(n_1, n_2 - i) - \zeta n_2} \\
&= \exp\left\{-\sum_{i=0}^{k} \log\left(1 - \frac{2\zeta n_2}{(n_2 - i)^2}\right)\right\} \\
&\approx \exp\left\{2\zeta n_2 \sum_{i=0}^{k} \frac{1}{(n_2 - i)^2}\right\} \qquad \text{if } \frac{2\zeta n_2}{(n_2 - i)^2} \approx 0 \\
&\approx \exp\left\{2\zeta \int_0^x \frac{1}{(1-y)^2}\, dy\right\} = \exp\left\{\frac{2\zeta x}{1-x}\right\}
\end{aligned} \tag{65}$$

if $\frac{k}{n_2} \to x \in (0,1)$ and $n_2 \to \infty$. Putting (65) and (60) into (64), we obtain the desired (62) and thus Lemma 1. □

We now return to the proof of Theorem 1. Lemma 1 implies that $(n_2\mathcal{T}_1,\, \mathcal{N}_2(\mathcal{T}_1)/n_2)$ converges in distribution to $(Z_1, 1-Y_1)$ as $n_2 \to \infty$. Since $Y_1 < 1$ almost surely, we have $\mathcal{N}_2(\mathcal{T}_1) \to \infty$ in the sense that

$$\lim_{n_2\to\infty} \mathbb{P}_{\mathbf{n}}(\mathcal{N}_2(\mathcal{T}_1) > M) = 1 \quad \text{for all } M \in (0,\infty). \tag{66}$$

As in (57), by definition, $\mathcal{T}_2$ is given by

$$\mathcal{T}_2 = \mathcal{T}_1 + \sum_{i=0}^{\sharp_2} \xi_i^{(2)},$$

where $\sharp_2$ is the number of downward jumps starting in state $(n_1-1, \mathcal{N}_2(\mathcal{T}_1))$ up to the second decrease in the first coordinate, i.e. to n_1-2. Like before, $\{\xi_i^{(2)}\}_{i=0}^{\sharp_2-1}$ are the times between these jumps, with $\xi_{\sharp_2}^{(2)}$ being the time for first coordinate to hit n_1-2 starting at the penultimate states $(n_1-1, \mathcal{N}_2(\mathcal{T}_1)-\sharp_2)$. As in (58), $\mathcal{N}_2(\mathcal{T}_2)$ is either $\mathcal{N}_2(\mathcal{T}_1)-\sharp_2$ or $\mathcal{N}_2(\mathcal{T}_1)-\sharp_2+1$.

As $n_2 \to \infty$, $\mathcal{N}_2(\mathcal{T}_1) \to \infty$ in the sense of (66). Hence the same argument that leads to Lemma 1 can be applied again, starting at the new location $(n_1-1, \mathcal{N}_2(\mathcal{T}_1))$. More precisely, by computing moment generating functions as before, and applying the strong Markov property of the random walk $\{(N_1(t), N_2(t))\}_{t\in\mathbb{R}_+}$ at the stopping time $\mathcal{T}_1$, we obtain the joint convergence

$$\left(n_2\sum_{i=0}^{\sharp}\xi_i,\, (n_2-\sharp)\sum_{i=0}^{\sharp_2}\xi_i^{(2)}\,;\, \frac{\sharp}{n_2}, \frac{\sharp_2}{n_2-\sharp}\right) \xrightarrow{\mathcal{L}} (Z_1, Z_2\,;\, Y_1, Y_2)$$

under $\mathbb{P}_{\mathbf{n}}$ as $n_2 \to \infty$, where $\{Z_1, Z_2, Y_1, Y_2\}$ are independent variables defined in Theorem 1. This implies the convergence in distribution

$$\left(n_2\mathcal{T}_1,\, n_2(\mathcal{T}_2-\mathcal{T}_1)\,;\, \frac{\mathcal{N}_2(\mathcal{T}_1)}{n_2}, \frac{\mathcal{N}_2(\mathcal{T}_2)}{n_2}\right)$$
$$\xrightarrow{\mathcal{L}} \left(Z_1, \frac{Z_2}{1-Y_1}\,;\, 1-Y_1,\, (1-Y_1)(1-Y_2)\right)$$

under $\mathbb{P}_{\mathbf{n}}$, as $n_2 \to \infty$. Continuing this way, by letting $\sharp_i$ be the number of downward jumps starting at $(n_1-i+1, \mathcal{N}_2(\mathcal{T}_{i-1}))$ before hitting the vertical line $\{(n_1-i, y) : y \in \mathbb{Z}_+\}$ for $i \geq 1$, we obtain the desired convergence in Theorem 1. □

A.2 Low-count branches of general coalescent trees

Here we prove the non-nestedness and Poisson-independence of low-count mutations, which we assumed in Section 2. We do this first for fixed trees then for the random, general coalescent trees of Griffiths and Tavaré (1998). We also present the computation of the probability generating function, $G_{n,k}(x,y)$, of the count of the variant of interest and its number of latent mutations.

A.2.1 Nested mutation on a fixed tree

Let $\mathbf{T}_n$ be a fixed (non random) tree with n leaves. We suppose the tree is ultrametric, that is the leaves have the same distance H_n from the root. We call H_n the height of $\mathbf{T}_n$. Consistent with the main text, we adopt the following notation for some relevant properties of $\mathbf{T}_n$, for the most part suppressing the dependence on n for simplicity:

1. T_k is the length of the time during which there are exactly k lineages ancestral to the sample, for $k \in \{2, 3, \cdots, n\}$.
2. τ_j for $j \in \{1, \cdots, n-1\}$, is the total length of branches in $\mathbf{T}_n$ that have j descendants. We suppose there are m_j such branches with lengths $\{\tau_{j,k}\}_{k=1}^{m_j}$. Then $\tau_j = \sum_{k=1}^{m_j} \tau_{j,k}$.
3. T_{total} is the total branch length, the sum of all the branches in $\mathbf{T}_n$, which is equal to $\sum_{k=2}^{n} k\, T_k = \sum_{j=1}^{n-1} \tau_j$.
4. For a positive integer b, we define a collection $\{\Gamma_i^{(b)}\}_{i=1}^{m_b}$ of disjoint connected subtrees of the coalescent tree as follows: Each of the m_b branches with b descendants in the sample (say the i-th one) subtends b leaves in the coalescent tree and gives rise to a subtree $\Gamma_i^{(b)}$ which contains that branch. We say **nested mutation up to count** b occurs on $\mathbf{T}_n$ if there exist two mutations on $\Gamma_i^{(b)} \subset \mathbf{T}_n$ for some $i \in \{1, 2, \cdots, m_b\}$. Figure 8 illustrates this for $b = 4$.

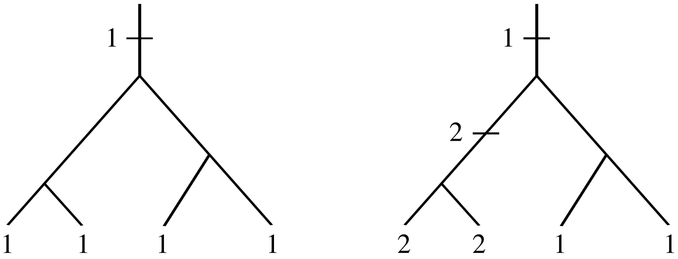


Figure 8: Two subtrees in $\{\Gamma_i^{(4)}\}$. The subtree on the left has one mutation which is labeled 1 and has count four. The subtree on the right has nested mutations, with the mutation labeled 1 in count two and another labeled 2 also in count two.

We assume that mutations arise as a Poisson point process on the tree with constant rate $\theta/2$ per unit length. Theorem 2 below holds for any fixed ultrametric tree (it can be binary or have multiple mergers, or even be a star tree).

Theorem 2 (Nested mutation on fixed trees). *Let $\mathbf{T}_n$ be a fixed ultrametric tree with n leaves. For any positive integer b and for any $\theta \in (0, \infty)$, the probability that nested mutation up to count b occurs is bounded above by*

$$\min\left\{\frac{\theta^2}{8} T_{\text{total}}\, H_n\,,\ \ \frac{\theta^2}{8}\, b^3 \sum_{j=1}^{b} \sum_{k=1}^{m_j} \tau_{j,k}^2\right\}. \tag{67}$$

In particular, the probability that nested mutation up to count b occurs tends to 0, as $n \to \infty$, if $\theta^2 \left(\max_{1 \le k \le m_j} \tau_{j,k}\right) \tau_j \to 0$ for $1 \le j \le b$.

Remark 3. There is good evidence that the upper bound bound $\frac{\theta^2}{8} T_{\text{total}}\, H_n$ is actually small for humans. For the gnomAD data we analyze in the main text, the expected number of mutations per site ($\theta T_{\text{total}}/2$) is between about 0.009 and 2.13. So $\theta T_{\text{total}}/2$ is not big with high probability. The rest of the upper bound, $\theta H_n/4$, should be proportional to the average pairwise difference per site (very nearly equal to this for random Kingman coalescent trees and large n) and this ranges from about 9×10^{-5} to about 0.02 for these same data. See Section 3.2.

Remark 4. The simpler bound $\frac{\theta^2}{8} T_{\text{total}}\, H_n$ can be weaker than the other bound $\frac{\theta^2}{8}\, b^3 \sum_{j=1}^{b} \sum_{k=1}^{m_j} \tau_{j,k}^2$ in (67) for large n. For the Kingman coalescent, $\mathbb{E}[T_{\text{total}}\, H_n] = O(\log n)$ is larger than $\mathbb{E}[\sum_{j=1}^{b} \sum_{k=1}^{m_j} \tau_{j,k}^2]$ since the latter tends to 0 as $n \to \infty$, by (70). For a star tree, however, both bounds are approximately $\theta^2 n\, H_n^2$ (up to a multiplicative constant).

Proof. The total number M_n of mutations on $\mathbf{T}_n$ is a Poisson variable with mean $c_n := \frac{\theta}{2} T_{\text{total}}$. Given the tree $\mathbf{T}_n$ and $M_n = k$, the k mutations are uniformly distributed on the tree. Hence the conditional probability that two given mutations are on the same subtree $\Gamma_i^{(b)}$ for some i is equal to

$$\sum_{i=1}^{m_b} \frac{|\Gamma_i^{(b)}|^2}{T_{\text{total}}^2},$$

where $|\Gamma_i^{(b)}|$ is the total branch lengths of the subtree $\Gamma_i^{(b)}$. Since there are $k(k-1)/2$ ways to choose two mutations out of k,

$$\begin{aligned}
&\mathbb{P}(\text{there are 2 mutations on } \Gamma_i^{(b)} \text{ for some } i \in \{1, 2, \cdots, m_b\}) \\
&\leq \sum_{k=0}^{\infty} e^{-c_n} \frac{c_n^k}{k!} \frac{k(k-1)}{2} \sum_{i=1}^{m_b} \frac{|\Gamma_i^{(b)}|^2}{T_{\text{total}}^2} \\
&= \frac{c_n^2}{2} \sum_{i=1}^{m_b} \frac{|\Gamma_i^{(b)}|^2}{T_{\text{total}}^2} \\
&= \frac{\theta^2}{8} \sum_{i=1}^{m_b} |\Gamma_i^{(b)}|^2. \qquad (68)
\end{aligned}$$

From here we can apply the simple bound $\sum_{i=1}^{m_b} |\Gamma_i^{(b)}|^2 \leq T_{\text{total}}\, H_n$ to obtain the first bound $\frac{\theta^2}{8} T_{\text{total}}\, H_n$ in (67). To get the second bound in (67), note that $|\Gamma_i^{(b)}| \leq b\, H_i^{(b)}$ for all $1 \leq i \leq m_b$, where $H_i^{(b)}$ is the height of the subtree $\Gamma_i^{(b)}$.

Furthermore, $H_i^{(b)}$ is the sum of at most b branch lengths, one from $\{\tau_{j,k}\}$ for $j = b, b-1, \cdots, 2, 1$, and these branches are pairwise disjoint for different i's (for $1 \leq i \leq m_b$). Hence

$$\sum_{i=1}^{m_b} |H_i^{(b)}|^2 \leq b \sum_{j=1}^{b} \sum_{k=1}^{m_j} \tau_{j,k}^2,$$

where we used the general inequality $|\sum_{k=1}^{b} a_k|^2 \leq b \sum_{k=1}^{b} a_k^2$. The bound in (67) now follows by putting these into (68).

□

A mutation on a tree (called a latent mutation in the main text) is said to **have count** j if the mutation is the most recent mutation in the lineages of exactly j individuals at the leaves of the tree; see Figure 8.

Theorem 3 (Poisson approximation for counts on a fixed tree). *Let $\mathbf{T}_n$ be a fixed coalescent tree with n leaves for $n \geq 2$. Let a_j be the number of mutations on $\mathbf{T}_n$ with counts j. If the probability that nested mutation up to count b occurs tends to 0 as $n \to \infty$, then for any positive integer b and any $\theta \in (0, \infty)$, the variables $\{a_j\}_{j=1}^{b}$ are asymptotically independent and $a_j \sim$ Poisson $\left(\frac{\theta}{2}\, \tau_j\right)$ for $1 \leq j \leq b$.*

Proof. If there is no nested mutation up to count b, then a_j is also equal to the number of mutations on the branches in $\mathbf{T}_n$ that have j descendants, for $1 \leq j \leq b$. Since these branches have total length τ_j and they are disjoint for different j's, the result follows from the assumption that mutations occur as a Poisson point process on the tree $\mathbf{T}_n$ with rate $\theta/2$. □

A.2.2 Nested mutation on random trees

We now suppose the tree $\mathbf{T}_n$ is a *random* binary tree (for $n \geq 2$), in particular the general coalescent tree of Griffiths and Tavaré (1998). For each $n \geq 2$, $\{T_k\}_{k=2}^n$ is a sequence of positive random variables representing the times during which there are k lineages in $\mathbf{T}_n$. The branching structure of $\mathbf{T}_n$ is independent of the times $\{T_k\}_{k=2}^n$. Looking forward in time, whenever there is a branching event, an existing lineage is chosen uniformly at random to split into two.

Following Griffiths and Tavaré (1998, eqn. (2.2)) we let $\lambda(t)$ be the the population size at time t in the past divided by the current population size. As in (45), $\lambda(t) = e^{-\alpha t}$ with $\alpha > 0$ corresponds to an exponentially growing population.

Theorem 4 (Nested mutation on random trees for fixed θ)**.** *Let $b \in \mathbb{N}$. Suppose for $1 \leq j \leq b$,*

$$\lim_{n\to\infty} \mathbf{E}_n \left[\sum_{k=1}^{m_j} \tau_{j,k}^2 \right] = 0, \tag{69}$$

where the expectation $\mathbf{E}_n$ averages over all realizations of $\mathbf{T}_n$. Then the probability that nested mutation up to count b occurs is bounded above by $C_{b,n}\,\theta^2$, where $\{C_{b,n}\}_{n\geq 2}$ are constants that tend to 0 as $n \to \infty$. Furthermore, (69) *holds for the generalized coalescent trees of Griffiths and Tavaré (1998) when $\sup_{t\geq 0} \lambda(t) < \infty$ (which includes any growing population).*

Proof. The first statement follows directly from Theorem 2. By the fact $\sum_{k=1}^{m_j} \tau_{j,k}^2 \leq \left(\max_{1\leq k\leq m_j} \tau_{j,k}\right) \tau_j$ and the Cauchy-Schwarz inequality, we have

$$\mathbf{E}_n \left[\sum_{k=1}^{m_j} \tau_{j,k}^2 \right] \leq \sqrt{\mathbf{E}_n[\tau_j^2]\, \mathbf{E}_n \left[\Big(\max_{1\leq k\leq m_j} \tau_{j,k} \Big)^2 \right]}. \tag{70}$$

Hence assumption (69) is satisfied if

$$\lim_{n\to\infty} \mathbf{E}_n \left[\Big(\max_{1\leq k\leq m_j} \tau_{j,k} \Big)^2 \right] = 0 \tag{71}$$

$$\limsup_{n\to\infty} \mathbf{E}_n \left[\tau_j^2 \right] < \infty, \tag{72}$$

for $1 \leq j \leq b$. The second statement now follows from Lemma 2, Lemma 3, and Proposition 1 below. □

Lemma 2 concerns assumption (71). For reference, we note that it is satisfied, and hence (71) is satisfied, if T_k are exponential variables with parameter λ_k where $\sum_{k=2}^{\infty} \frac{1}{\lambda_k} < \infty$. This is true for the Kingman coalescent which has $\lambda_k = k(k-1)/2$.

Lemma 2. *Suppose* $\limsup_{n\to\infty} \sum_{k=2}^{n} T_k$ *has finite p-th moment, where $p > 0$. Then* $\max_{1\leq k\leq m_j} \tau_{j,k} \to 0$ *in L^p, as $n \to \infty$.*

Proof. Consider the random tree $\mathbf{T}_n$ and recall that T_k is the length of the time during which there are exactly k lineages ancestral to the sample in $\mathbf{T}_n$. These k lineages are segments of length T_k of the branches of the genealogy, and each of them is called a line of state k.

Let $A_\ell^{(k,n)}$ be the number of descendants in $\mathbf{T}_n$ of the ℓ-th line of state k. Note that $A_\ell^{(k,n)} \geq 1$ for $\ell \in \{1, 2, \cdots, k\}$, and $\sum_{\ell=1}^k A_\ell^{(k,n)} = n$. Since the branching structure is independent of $\{T_k\}_{k=2}^n$, we can assume without loss of generality that $\{A_\ell^{(k,n)} : 1 \leq \ell \leq k,\, 2 \leq k \leq n,\, n \geq 2\}$ are all defined on the same probability space. By exchangeability—in particular see Bertoin (2006, Proposition 2.8)—the random vector $\frac{1}{n}(A_1^{(k,n)}, \cdots, A_k^{(k,n)})$ converges almost surely to a random vector that has the symmetric Dirichlet distribution on the simplex $\{(x_i)_{i=1}^k \in \mathbb{R}_+^k :\, x_1 + \cdots + x_k = n\}$. Therefore, with probability one,

$$\lim_{n\to\infty} A_\ell^{(k,n)} = +\infty \quad \text{for all } k \geq 1 \text{ and } \ell = 1, 2, \cdots, k. \tag{73}$$

Since $\sum_{k=2}^\infty T_k$ is finite almost surely, the trees $\{\mathbf{T}_n\}_{n\geq 2}$ have uniformly bounded height almost surely. So (73) implies that with probability one,

$$\limsup_{n\to\infty} \sup_{1\leq k\leq m_j} \tau_{jk} = 0 \quad \text{for all } j \geq 1.$$

Since $\max_{1\leq k\leq m_j} \tau_{j,k} < \sum_{k=2}^n T_k$, by the assumption on $\{T_k\}$ and the Dominated Convergence Theorem, $\max_{1\leq k\leq m_j} \tau_{j,k} \to 0$ in L^p as $n \to \infty$. □

Next consider assumption (72). For the Kingman coalescent, τ_j is close to its mean $\mathbf{E}_n[\tau_j] = 2/j$ because for n large enough,

$$\mathrm{Var}(\tau_j) = 4\sigma_{jj} \leq \frac{4(j+1)\log n}{n}, \tag{74}$$

where σ_{jj} is defined in Fu (1995, eqns. (1)-(2)). This follows from the fact $\beta_n(j) \approx \frac{2\log n}{n}$ as $n \to \infty$ for each $j \geq 1$ (Fu, 1995, eqn. (5)). Hence

$$\limsup_{n\to\infty} \mathbf{E}_n[\tau_j^2] \leq \left(\frac{2}{j}\right)^2.$$

Lemma 3. *Suppose there exists a constant $C_* \in (0, \infty)$ such that*

$$\sup_{n\geq 2} \mathbf{E}_n[T_k^2] \leq \frac{C_*}{k^4} \qquad \textit{for all } k \geq 2. \tag{75}$$

Then $\mathbf{E}_n[\tau_j] \leq \frac{\sqrt{C_*}}{j}$ for all $j \geq 1$ and $\limsup_{n\to\infty} \mathbf{E}_n\left[\tau_j^2\right] < \infty$.

Proof. For realized values of T_k, the argument in Fu (1995, page 181) gives

$$\tau_j = \sum_{k=2}^n \sum_{\ell=1}^k \epsilon_{k,\ell}(j)\, T_k = \sum_{k=2}^n T_k \sum_{\ell=1}^k \epsilon_{k\ell}(j),$$

where $\epsilon_{k\ell}(j) = 1_{\{A_\ell^{(k,n)}=j\}}$ is the indicator variable, where $A_\ell^{(k,n)}$ is the number of descendants in $\mathbf{T}_n$ of the ℓ-th line of state k defined in the proof of Lemma 2.

Using the independence between $\{T_k\}_{k\geq 2}$ and the branching structure, and following the notation in Fu (1995, eqns. (18)-(19)), the conditional expectation of τ_j, given $\{T_k\}_{k=2}^n$, is

$$\mathbf{E}_n[\tau_j \mid \{T_k\}_{k=2}^n] = \sum_{k=2}^n T_k\, k\, p(k,j) \tag{76}$$

and that of τ_j^2, given $\{T_k\}_{k=2}^n$, is

$$\begin{aligned}\mathbf{E}_n[\tau_j^2 \mid \{T_k\}_{k=2}^n] = &\sum_{k=2}^n T_k^2 \left(kp(k,j) + k(k-1)p(k,j;k,j)\right)\\ &+ 2\sum_{k<k'} T_k T_{k'}\, kk'\, p(k,j;k',j),\end{aligned} \tag{77}$$

where the deterministic functions $p(k,j)$, $p(k,j;k',j)$ do not depend on $\{T_k\}$. From Fu (1995),

$$p(k,j) = \frac{\binom{n-j-1}{k-2}}{\binom{n-1}{k-1}} = \frac{\binom{n-k}{j-1}}{\binom{n-1}{j}}\,\frac{k-1}{j}, \qquad p(k,j;k,j) = \frac{\binom{n-2j-1}{k-3}}{\binom{n-1}{k-1}}$$

and for $2 \leq k < k' \leq n$,

$$\begin{aligned}p(k,j;k',j) =&\frac{k-1}{k'(k'-1)}\, p(k',j)\\ &+ \sum_t \frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}}\,\frac{(k-1)(k'-t)}{tk'}\,\frac{\binom{j-1}{t-1}\binom{n-2j-1}{k'-2-t}}{\binom{n-1}{k'-1}},\end{aligned}$$

where the sum is taken over $1 \leq t \leq \min\{j,\, k'-2,\, k'-k+1\}$.

The first and the second moments of τ_i are obtained averaging over $\{T_k\}_{k=2}^n$ in (76) and (77). The bound $\mathbf{E}_n\left[\tau_j\right] \leq \frac{\sqrt{C_*}}{j}$ follows from the same calculation in Fu (1995, eqn. (22)). By (77), the fact $\mathbf{E}_n[T_k T_{k'}] \leq (\mathbf{E}_n[T_k^2]\,\mathbf{E}_n[T_{k'}^2])^{1/2}$ and assumption (75), $\limsup_{n\to\infty} \mathbf{E}_n\left[\tau_j^2\right] < \infty$ holds also for our random trees. □

Remark 5. As in Theorem 2, we can use an alternate assumption than 69. For any positive integer b, the probability that nested mutation up to count b occurs is bounded above by $\frac{\theta^2}{8}\,\mathbf{E}_n\left[T_{\text{total}}\, H_n\right]$ which tends to 0 if $\theta^2\,\mathbf{E}_n\left[T_{\text{total}}\, H_n\right] \to 0$. For Kingman coalescent trees, this would require that $\theta \to 0$.

We now check that the assumption (69) in Theorem 4 holds for the generalized coalescent tree of Griffiths and Tavaré (1998).

Proposition 1. *Suppose* $C_0 := \sup_{t\geq 0}\lambda(t) < \infty$*. Then* $\{T_k:\ 2 \leq k \leq n,\, n \geq 2\}$ *satisfy the conditions in both Lemma 2 (with* $p = 2$*) and Lemma 3. In particular, (69) is satisfied and so the conclusion of Theorem 4 holds.*

Proof. The joint distribution of $\{T_k\}_{k=2}^n$ is determined by the function λ; see Griffiths and Tavaré (1994b). We can construct $\{T_k\}_{k=2}^n$ in terms of λ as follows: let $\{D_n(t)\}_{t\in\mathbb{R}_+}$ be a pure death process with rate $\binom{k}{2}$ at state $k \in \{1, 2, \cdots, n\}$, starting at $D_n(0) = n$, and let

$$D_n^{(\lambda)}(t) = D_n\left(\int_0^t \frac{1}{\lambda(u)}\, du\right) \tag{78}$$

be a time-changed pure death process. Then

$$T_k = \int_0^\infty \mathbf{1}\{D_n^{(\lambda)}(t) = k\}dt = \sigma_{n-k+1} - \sigma_{n-k},$$

for $2 \leq k \leq n$, where $\sigma_1 < \sigma_2 < \cdots < \sigma_{n-1}$ are the jump times of $D_n^{(\lambda)}$ (by convention $\sigma_0 = 0$).

By (78), the jump times of the pure death process D_n, denoted by $\widetilde{\sigma}_1 < \widetilde{\sigma}_2 < \cdots < \widetilde{\sigma}_{n-1}$, are given by $\int_0^{\sigma_j} \frac{1}{\lambda} = \widetilde{\sigma}_j$ for $1 \leq j \leq n-1$. Hence, with the convention $\sigma_0 = 0$, for $0 \leq j \leq n-2$ we have

$$\frac{\sigma_{j+1} - \sigma_j}{C_0} \leq \int_{\sigma_j}^{\sigma_{j+1}} \frac{1}{\lambda(t)}\, dt = \widetilde{\sigma}_{j+1} - \widetilde{\sigma}_j.$$

These give $T_k = \sigma_{n-k+1} - \sigma_{n-k} \leq (\widetilde{\sigma}_{n-k+1} - \widetilde{\sigma}_{n-k})\, C_0$ for all $2 \leq k \leq n$.

Since $\widetilde{\sigma}_{n-k+1} - \widetilde{\sigma}_{n-k}$ is equal in distribution to the analogue of T_k for the Kingman coalescent, T_k is stochastically dominated by C_0 times an exponential variable with parameter $k(k-1)/2$ for all $2 \leq k \leq n$. The desired statement now follows since (71) and (72) are satisfied. □

A.2.3 Replacing τ_j by its mean

By using the expected coalescence times denoted $\bar{\tau}_i$ in the main tex, we implicitly assumed that different sites have different trees and that these are all drawn from the same distribution. Theorem 5 below asserts that even though the mutant counts at each site are conditional on the realization of the tree at that site, we can replace τ_j by its expectation $\mathbf{E}_n[\tau_j]$ in Theorem 3 when the trees are random and satisfy suitable assumptions. The key reason is that τ_j is close to its mean, as made precise in Lemma 4.

Lemma 4. *Suppose* (75) *holds and that the covariance*

$$\mathrm{Cov}(T_k, T_{k'}) \leq \frac{C_n}{k(k-1)k'(k'-1)} \tag{79}$$

for $2 \leq k < k' \leq n$ *and* $n \geq 2$*, where* $\{C_n\}$ *is a sequence that tends to 0 as* $n \to \infty$*. Then for each* $j \geq 1$*, the variance* $\mathrm{Var}(\tau_j) \to 0$ *as* $n \to \infty$*. In particular,* $|\tau_j - \mathbb{E}[\tau_j]\,| \to 0$ *in* $L^2(\mathbb{P})$ *as* $n \to \infty$*.*

Proof. By further taking expectations in (76) and (77) with respect to $\mathbf{E}_n$, we obtain the variance

$$\begin{aligned}\mathrm{Var}(\tau_j) =& \mathbf{E}_n[\tau_j^2] - (\mathbf{E}_n[\tau_j])^2 \\ =& 2\sum_{k<k'} kk' \Big(\mathbf{E}_n[T_k T_{k'}]\, p(k,j;k',j) \\ & \qquad - \mathbf{E}_n[T_k]\mathbf{E}_n[T_{k'}]\, p(k,j)p(k',j)\Big)\end{aligned} \tag{80}$$

up to an $O\left(\frac{\log n}{n}\right)$ term. This follows from Fu (1995, eqns. (24)-(25)) and assumption (75) in Lemma 3. This also leads to (74).

By assumptions (75) and (79), the double sum in (80) is bounded above by

$$C_n \sum_{k<k'} \frac{p(k,j;k',j)}{(k-1)(k'-1)} + C_* \sum_{k<k'} \frac{p(k,j;k',j) \; - \; p(k,j)p(k',j)}{(k-1)(k'-1)}. \tag{81}$$

By Fu (1995, eqns. (29) and (22)), the first and second terms of (81) are of order $o(n)$ and $O\left(\frac{\log n}{n}\right)$, respectively, as $n \to \infty$ for each $j \geq 1$. The completes the proof of $\lim_{n\to\infty} \mathrm{Var}(\tau_j) = 0$. The latter implies, by Chebyshev's inequality, that $\tau_j - \mathbb{E}[\tau_j] \to 0$ in L^2 as $n \to \infty$. □

Theorem 5 (Poisson approximation for counts across loci). *Let $\{\mathbf{T}_n\}_{n\geq 2}$ be a sequence of random coalescent trees which are the generalized coalescent trees of Griffiths and Tavaré (1998). Suppose $\sup_{t\geq 0} \lambda(t) < \infty$ and assumption (79) holds. Let a_j be the number of mutations on $\mathbf{T}_n$ with counts j. Then for any positive integer b and any $\theta \in (0, \infty)$, the variables $\{a_j\}_{j=1}^{b}$ are asymptotically independent and $a_j \sim$ Poisson $\left(\frac{\theta}{2}\mathbf{E}_n[\tau_j]\right)$ for $1 \leq j \leq b$, as $n \to \infty$.*

Proof. By Theorem 4, the probability that nested mutation up to count b occurs tends to 0 as $n \to \infty$. The result then follows from Lemma 4 and Theorem 3. □

It can be checked that exponentially growing popolations clearly satisfy $\sup_{t\geq 0} \lambda(t) < \infty$ and also assumption (79). The conclusions of Theorems 4 and 5 then hold for the generalized coalescent trees of Griffiths and Tavaré (1998) when $\lambda(t) = e^{\alpha t}$ for $t \in \mathbb{R}_+$ for some $\alpha > 0$.

Equipped with Theorem 5, we write $\bar{\tau}_i = \mathbf{E}_n[\tau_i]$ as in the main text and compute the probability generating function $G_{n,k}$ of the count of the variant of interest and its number of latent mutations. The count of the variant of interest is $n = \sum_i i a_i$ and its number of latent mutations is $k = \sum_i a_i$. Hence

$$\begin{aligned}
G_{n,k}(x, y) &= \sum_{(a_1,a_2,\ldots)} P(a_1, a_2, \ldots) x^n y^k \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \sum_{(a_1,a_2,\ldots)} x^{\sum i a_i} y^{\sum a_i} \prod_{i\geq 1} \frac{(\theta\bar{\tau}_i/2)^{a_i}}{a_i!} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \sum_{(a_1,a_2,\ldots)} \prod_{i\geq 1} \frac{x^{i a_i} y^{a_i} (\theta\bar{\tau}_i/2)^{a_i}}{a_i!} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \prod_{i\geq 1} \sum_{a_i \geq 0} \frac{x^{i a_i} y^{a_i} (\theta\bar{\tau}_i/2)^{a_i}}{a_i!} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \prod_{i\geq 1} e^{x^i y \theta \bar{\tau}_i/2} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} e^{\frac{\theta}{2} y \sum_i x^i \bar{\tau}_i} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \sum_{k=0}^{\infty} \frac{(\frac{\theta}{2})^k y^k}{k!} \left(\sum_i x^i \bar{\tau}_i\right)^k \qquad (82)
\end{aligned}$$

as declared in the main text.

Figure 9: Plots like those in Figure 7 for each of the 97 mutation-rate bins. (continues on next page)

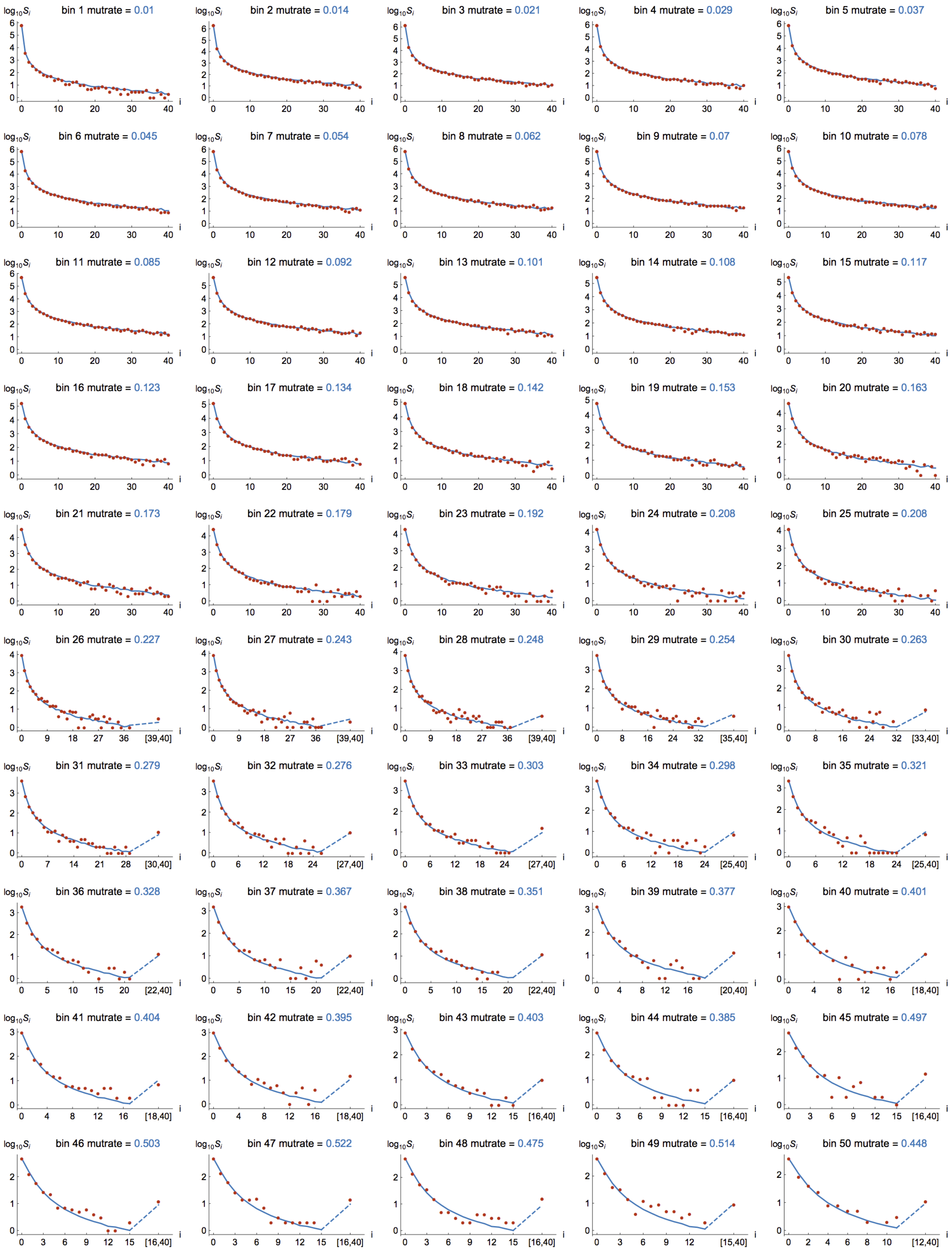
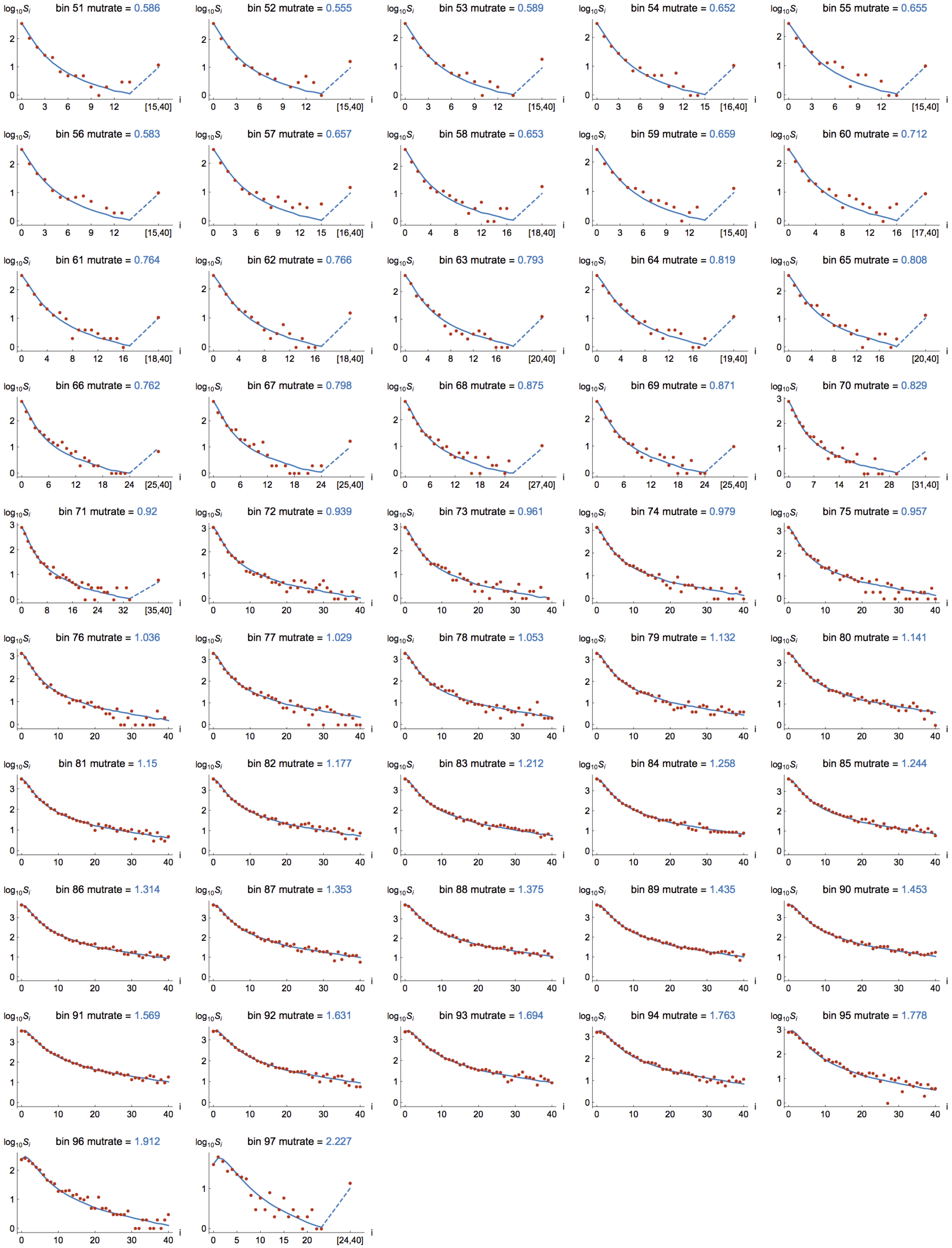


Figure 9: (continued from previous page)



References

Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions.* Dover, New York, 1964.

Guillaume Achaz. Frequency spectrum neutrality tests: One for all and all for one. *Genetics*, 183 (1):249–258, 2009. doi: 10.1534/genetics.109.104042.

Varun Aggarwala and Benjamin F Voight. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48(4):349–355, 2016. doi: 10.1038/ng.3511.

Richard Arratia, A D Barbour, and Simon Tavaré. Poisson process approximations for the Ewens sampling formula. *The Annals of Applied Probability*, 2(3):519 – 535, 1992. doi: 10.1214/aoap/1177005647.

Richard Arratia, A D Barbour, and Simon Tavaré. *Logarithmic Combinatorial Structures: A Probabilistic Approach.* EMS monographs in mathematics. European Mathematical Society, 2003.

Richard Arratia, A D Barbour, and Simon Tavaré. Exploiting the Feller coupling for the Ewens sampling formula. *Statistical Science*, 31(1):27 – 29, 2016. doi: 10.1214/15-STS537.

Ellen Baake and Robert Bialowons. Ancestral processes with selection: branching and Moran models. *Banach Center Publications*, 80:33–52, 2008. doi: 10.4064/bc80-0-2.

Jean Bertoin. *Random Fragmentation and Coagulation Processes.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2006. doi: 10.1017/CBO9780511617768.

Anand Bhaskar and Yun S Song. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *The Annals of Statistics*, 42(6):2469 – 2493, 2014. doi: 10.1214/14-AOS1264.

Anand Bhaskar, John A Kamm, and Yun S Song. Approximate sampling formulae for general finite-alleles models of mutation. *Advances in Applied Probability*, 44(2):408–428, 2012. doi: 10.1239/aap/1339878718.

Adrian P Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504, 1980. doi: 10.1093/nar/8.7.1499.

John M Braverman, Richard R Hudson, Norman L Kaplan, Charles H Langley, and Wolfgang Stephan. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140(2):783–796, 1995. doi: 10.1093/genetics/140.2.783.

Conrad J Burden and Robert C Griffiths. The stationary distribution of a sample from the Wright-Fisher diffusion model with general small mutation rates. *Journal of Mathematical Biology*, 78(4): 1211–1224, 2019. doi: 10.1007/s00285-018-1306-y.

Conrad J Burden and Yurong Tang. An approximate stationary solution for multi-allele neutral diffusion with low mutation rates. *Theoretical Population Biology*, 112:22–32, 2016. doi: 10.1016/j.tpb.2016.07.005.

Conrad J Burden and Yurong Tang. Rate matrix estimation from site frequency data. *Theoretical Population Biology*, 113:23–33, 2017. doi: 10.1016/j.tpb.2016.10.001.

Carlos D Bustamante, John Wakeley, Stanley Sawyer, and Daniel L Hartl. Directional selection and the site-frequency spectrum. *Genetics*, 159(4):1779–1788, 2001. doi: 10.1093/genetics/159.4.1779.

Nicolas Champagnat and Amaury Lambert. Splitting trees with neutral Poissonian mutations I: Small families. *Stochastic Processes and their Applications*, 122(3):1003–1033, 2012. doi: 10.1016/j.spa.2011.11.002.

Nicolas Champagnat and Amaury Lambert. Splitting trees with neutral poissonian mutations ii: Largest and oldest families. *Stochastic Processes and their Applications*, 123(4):1368–1414, 2013. doi: 10.1016/j.spa.2012.11.013.

David Cheek and Tibor Antal. Mutation frequencies in a birth-death branching process. *The Annals of Applied Probability*, 28(6):3922 – 3947, 2018. doi: 10.1214/18-AAP1413.

Hua Chen and Kun Chen. Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. *Genetics*, 194(3):721–736, 2013. doi: 10.1534/genetics.113.151522.

Fausto F Crespo, David Posada, and Carsten Wiuf. Coalescent models derived from birth-death processes. *Theoretical Population Biology*, 142:1–11, 2021. doi: https://doi.org/10.1016/j.tpb.2021.09.003.

Maria De Iorio and Robert C Griffiths. Importance sampling on coalescent histories. I. *Advances in Applied Probability*, 36(2):417–433, 2004. doi: 10.1239/aap/1086957579.

Michael M Desai and Joshua B Plotkin. The polymorphism frequency spectrum of finitely many sites under selection. *Genetics*, 180(4):2175–2191, 2008. doi: 10.1534/genetics.108.087361.

Peter Donnelly and Simon Tavaré. The population genealogy of the infinitely-many neutral alleles model. *Journal of Mathematical Biology*, 25(4):381–391, 1987. doi: 10.1007/BF00277163.

Karin S Dorman, Janet S Sinsheimer, and Kenneth Lange. In the garden of branching processes. *SIAM Review*, 46(2):202–229, 2004. doi: 10.1137/S0036144502417843.

Richard Durrett. *Branching process models of cancer*, volume 1.1 of *Mathematical Biosciences Institute Lecture Series*. Springer Cham, Heidelberg New York Dordrecht London, 2015. doi: 10.1007/978-3-319-16065-8.

Rick Durrett. Population genetics of neutral mutations in exponentially growing cancer cell populations. *The Annals of Applied Probability*, 23(1):230 – 250, 2013. doi: 10.1214/11-AAP824.

Bjarki Eldon, Matthias Birkner, Jochen Blath, and Fabian Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3): 841–856, 01 2015. doi: 10.1534/genetics.114.173807.

Warren J Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87–112, 1972. doi: 10.1016/0040-5809(72)90035-4.

Warren J Ewens. A note on the sampling theory for infinite alleles and infinite sites models. *Theoretical Population Biology*, 6(2):143–148, 1974. doi: 10.1016/0040-5809(74)90020-3.

Warren J Ewens. *Mathematical Population Genetics*. Springer-Verlag, Berlin, 1979.

Warren J Ewens. *Mathematical Population Genetics, Volume I: Theoretical Foundations*. Springer-Verlag, Berlin, 2004.

Paul Fearnhead. The common ancestor at a nonneutral locus. *Journal of Applied Probability*, 39(1): 38–54, 2002. doi: 10.1239/jap/1019737986.

Luca Ferretti, Alice Ledda, Thomas Wiehe, Guillaume Achaz, and Sebastian E Ramos-Onsins. Decomposing the site frequency spectrum: The impact of tree topology on neutrality tests. *Genetics*, 207(1):229–240, 2017. doi: 10.1534/genetics.116.188763.

R A Fisher. The possible modification of the response of the wild type to recurrent mutations. *The American Naturalist*, 62(679):115–126, 1928. doi: 10.1086/280193.

Ronald A Fisher. The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh*, 50:205–220, 1930a. URL `https://hdl.handle.net/2440/15106`.

Ronald A Fisher. *The Genetical Theory of Natural Selection*. Clarendon, Oxford, 1930b.

Ronald A Fisher. A theoretical distribution for the apparent abundance of different species. *Journal of Animal Ecology*, 12:54–57, 1943. URL `https://hdl.handle.net/2440/15246`.

Yun-Xin Fu. Statistical properties of segregating sites. *Theoretical Population Biology*, 48(2): 172–197, 1995. doi: 10.1006/tpbi.1995.1025.

Feng Gao and Alon Keinan. Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. *Genetics*, 202(1):235–245, 2016. doi: 10.1534/genetics.115.180570.

Elodie Gazave, Li Ma, Diana Chang, Alex Coventry, Feng Gao, Donna Muzny, Eric Boerwinkle, Richard . Gibbs, Charles F Sing, Andrew G Clark, and Alon Keinan. Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences*, 111(2):757–762, 2014. doi: 10.1073/pnas.1310398110.

Nick Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Research*, 21(10):2487–2491, 1993. doi: 10.1093/nar/21.10.2487.

Robert C Griffiths and Simon Tavaré. Ancestral inference in population genetics. *Statistical Science*, 9(3):307 – 319, 1994a. doi: 10.1214/ss/1177010378.

Robert C Griffiths and Simon Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344 (1310):403 – 410, 1994b. doi: 10.1098/rstb.1994.0079.

Robert C Griffiths and Simon Tavaré. The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models*, 14(1-2):273–295, 1998. doi: 10.1080/15326349808807471.

Einar Bjarki Gunnarsson, Kevin Leder, and Jasmine Foo. Exact site frequency spectra of neutrally evolving tumors: A transition between power laws reveals a signature of cell viability. *Theoretical Population Biology*, 142:67–90, 2021. doi: https://doi.org/10.1016/j.tpb.2021.09.004.

Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genetics*, 5(10):1–11, 10 2009. doi: 10.1371/journal.pgen.1000695.

J B S Haldane. The part played by recurrent mutation in evolution. *The American Naturalist*, 67 (708):5–19, 1933. doi: 10.1086/280465.

Arbel Harpak, Anand Bhaskar, and Jonathan K Pritchard. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS Genetics*, 12:e1006489, 2016. doi: 10.1371/journal.pgen.1006489.

Asger Hobolth and Carsten Wiuf. The genealogy, site frequency spectrum and ages of two nested mutant alleles. *Theoretical Population Biology*, 75(4):260–265, 2009. doi: 10.1016/j.tpb.2009.02.001. Sam Karlin: Special Issue.

Richard R Hudson. Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37(1):203–217, 1983. doi: 10.1111/j.1558-5646.1983.tb05528.x.

Paul A Jenkins and Yun S Song. The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theoretical Population Biology*, 80(2):158–173, 2011. doi: 10.1016/j.tpb.2011.04.001.

Paul A Jenkins, Jonas W Mueller, and Yun S Song. General triallelic frequency spectrum under demographic models with variable population size. *Genetics*, 196(1):295–311, 2014. doi: 10.1534/genetics.113.158584.

Kelsey E Johnson and Benjamin F Voight. Identifying non-identical-by-descent rare variants in population-scale whole genome sequencing data. *bioRxiv*, 2020. doi: 10.1101/2020.05.26.117358.

Ingemar Kaj and Carina F Mugal. The non-equilibrium allele frequency spectrum in a poisson random field framework. *Theoretical Population Biology*, 111:51–64, 2016. doi: 10.1016/j.tpb.2016.06.003.

Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, Laura D Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M England, Eleanor G Seaby, Jack A Kosmicki, Raymond K Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X Chong, Kaitlin E Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S Ware, Christopher Vittal, Irina M Armean, Louis Bergelson, Kristian Cibulskis, Kristen M Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E Talkowski, Carlos A Aguilar Salinas, Tariq Ahmad, Christine M Albert, Diego Ardissino, Gil Atzmon, John Barnard, Laurent Beaugerie, Emelia J Benjamin, Michael Boehnke, Lori L Bonnycastle, Erwin P Bottinger, Donald W Bowden, Matthew J Bown, John C Chambers, Juliana C Chan, Daniel Chasman, Judy Cho, Mina K Chung, Bruce Cohen, Adolfo Correa, Dana Dabelea, Mark J Daly, Dawood Darbar, Ravindranath Duggirala, Josée Dupuis, Patrick T Ellinor, Roberto Elosua, Jeanette Erdmann, Tõnu Esko, Martti Färkkilä, Jose Florez, Andre Franke, Gad Getz, Benjamin Glaser, Stephen J Glatt, David Goldstein, Clicerio Gonzalez, Leif Groop, Christopher Haiman, Craig Hanis, Matthew Harms, Mikko Hiltunen, Matti M Holi, Christina M Hultman, Mikko Kallela, Jaakko Kaprio, Sekar Kathiresan, Bong-Jo Kim, Young Jin Kim, George Kirov, Jaspal Kooner, Seppo Koskinen, Harlan M Krumholz, Subra Kugathasan, Soo Heon Kwak, Markku Laakso, Terho Lehtimäki, Ruth J F Loos, Steven A Lubitz, Ronald C W Ma, Daniel G MacArthur, Jaume Marrugat, Kari M Mattila, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, James B Meigs, Olle Melander, Andres Metspalu, Benjamin M Neale, Peter M Nilsson, Michael C O'Donovan, Dost Ongur, Lorena Orozco, Michael J Owen, Colin N A Palmer, Aarno Palotie, Kyong Soo Park, Carlos Pato, Ann E Pulver, Nazneen Rahman, Anne M Remes, John D Rioux, Samuli Ripatti, Dan M Roden, Danish Saleheen, Veikko Salomaa, Nilesh J Samani, Jeremiah Scharf, Heribert Schunkert, Moore B Shoemaker, Pamela Sklar, Hilkka Soininen, Harry Sokol, Tim Spector, Patrick F Sullivan, Jaana Suvisaari, E Shyong Tai, Yik Ying Teo, Tuomi Tiinamaija, Ming Tsuang, Dan Turner, Teresa Tusie-Luna, Erkki Vartiainen, Marquis P Vawter, James S Ware, Hugh Watkins, Rinse K Weersma, Maija Wessman, James G Wilson, Ramnik J Xavier, Benjamin M Neale, Mark J Daly, Daniel G MacArthur, and Genome Aggregation Database Consortium. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581:434–443, 2020. doi: 10.1038/s41586-020-2308-7.

Alon Keinan and Andrew G Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, 2012. doi: 10.1126/science.1217283.

Andrew D Kern and Jody Hey. Exact calculation of the joint allele frequency spectrum for isolation with migration models. *Genetics*, 207(1):241–253, 2017. doi: 10.1534/genetics.116.194019.

David A Kessler and Herbert Levine. Large population solution of the stochastic luria & delbr uck evolution model. *Proceedings of the National Academy of Sciences*, 110(29):11682–11687, 2013. doi: 10.1073/pnas.1309667110.

Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. *Genetics*, 61(4):893–903, 1969. URL `https://www.genetics.org/content/61/4/893`.

Motoo Kimura. Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology*, 2(2):174–208, 1971. doi: 10.1016/0040-5809(71)90014-1.

J F C Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43, 1982. doi: 10.2307/3213548.

Stephen M Krone and Claudia Neuhauser. Ancestral processes with selection. *Theoretical Population Biology*, 51(3):210–237, 1997. doi: 10.1006/tpbi.1997.1299.

Amaury Lambert. Species abundance distributions in neutral models with immigration or mutation and general lifetimes. *Journal of Mathematical Biology*, 63(1):57–72, 2011. doi: 10.1007/s00285-010-0361-9.

Kenneth Lange and Ru zong Fan. Branching process models for mutant genes in nonstationary populations. *Theoretical Population Biology*, 51(2):118–133, 1997. doi: 10.1006/tpbi.1997.1297.

Marguerite Lapierre, Amaury Lambert, and Guillaume Achaz. Accuracy of demographic inferences from the site frequency spectrum: The case of the Yoruba population. *Genetics*, 206(1):439–449, 2017. doi: 10.1534/genetics.116.192708.

D E Lea and A Coulson. The distribution of the numbers of mutants in bacterial populations. *Journal of Genetics*, 49(3):264–285, 1949. URL `https://www.ias.ac.in/article/fulltext/jgen/049/03/0264-0285`.

Ellen M Leffler, Kevin Bullaughey, Daniel R Matute, Wynn K. Meyer, Laure Séurel, Aarti Venkat, Peter Andolfatto, and Molly Przeworski. Revisiting an old riddle: What determines genetic diversity levels within species? *PLOS Biology*, 10(9):1–9, 2012. doi: 10.1371/journal.pbio.1001388.

Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A Kosmicki, Laramie E Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M Peloso, Ryan Poplin, Manuel A Rivas, Valentin Ruano-Rubio, Samuel A Rose, Douglas M Ruderfer, Khalid Shakir, Peter D Stenson, Christine Stevens, Brett P Thomas, Grace Tiao, Maria T Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C Florez, Stacey B Gabriel, Gad Getz, Stephen J Glatt, Christina M Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M Neale, Aarno Palotie, Shaun M Purcell, Danish Saleheen, Jeremiah M Scharf, Pamela Sklar, Patrick F Sullivan, Jaakko Tuomilehto, Ming T Tsuang, Hugh C Watkins, James G Wilson, Mark J Daly, Daniel G MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016. doi: 10.1038/nature19057.

Xiaoming Liu and Yun-Xin Fu. Exploring population size changes using snp frequency spectra. *Nature Genetics*, 47(5):555–559, 2015. doi: 10.1038/ng.3254.

S E Luria and M Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, 1943. doi: 10.1093/genetics/28.6.491.

P A P Moran. Random processes in genetics. *Proc. Camb. Phil. Soc.*, 54(1):60–71, 1958. doi: 10.1017/S0305004100033193.

P A P Moran. *Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford, 1962.

Rebekka Müller, Ingemar Kaj, and Carina F Mugal. A nearly neutral model of molecular signatures of natural selection after change in population size. *Genome Biology and Evolution*, 14(5), 2022. doi: 10.1093/gbe/evac058.

Simon Myers, Charles Fefferman, and Nick Patterson. Can one learn history from the allelic spectrum? *Theoretical Population Biology*, 73(3):342–348, 2008. doi: https://doi.org/10.1016/j.tpb.2008.01.001.

Claudia Neuhauser and Stephen M Krone. The genealogy of samples in models with selection. *Genetics*, 145(2):519–534, 1997. doi: 10.1093/genetics/145.2.519.

Hisashi Ohtsuki and Hideki Innan. Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. *Theoretical Population Biology*, 117:43–50, 2017. doi: https://doi.org/10.1016/j.tpb.2017.08.006.

Andrzej Polanski and Marek Kimmel. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1):427–436, 2003. doi: 10.1093/genetics/165.1.427.

Andrzej Polanski, Agnieszka Szczesna, Mateusz Garbulowski, and Marek Kimmel. Coalescence computations for large samples drawn from populations of time-varying sizes. *PLOS ONE*, 12(2): 1–22, 2017. doi: 10.1371/journal.pone.0170701.

Gladys Y P Poon, Caroline J Watson, Daniel S Fisher, and Jamie R Blundell. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. *Nature Genetics*, 53 (11):1597–1605, 2021. doi: 10.1038/s41588-021-00957-1.

Bruce Rannala and Montgomery Slatkin. Estimating the age of alleles by use of intraallelic variability. *The American Journal of Human Genetics*, 60(2):447–458, 1997. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1712388/`.

Zvi Rosen, Anand Bhaskar, Sebastien Roch, and Yun S Song. Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*, 210(2):665–682, 2018. doi: 10.1534/genetics.118.300733.

Ori Sargsyan. *Analytical and Simulation Results for the General Coalescent*. PhD thesis, University of Southern California, 2006.

Ori Sargsyan. An analytical framework in the general coalescent tree setting for analyzing polymorphisms created by two mutations. *Journal of Mathematical Biology*, 70(4):913–956, 2015. doi: 10.1007/s00285-014-0785-8.

Dominik Schrempf and Asger Hobolth. An alternative derivation of the stationary distribution of the multivariate neutral Wright-Fisher model for low mutation rates with a view to mutation rate estimation from site frequency data. *Theoretical Population Biology*, 114:88–94, 2017. doi: https://doi.org/10.1016/j.tpb.2016.12.001.

Vladimir B Seplyarskiy, Ruslan A Soldatov, Evan Koch, Ryan J McGinty, Jakob M Goldmann, Ryan D Hernandez, Kathleen Barnes, Adolfo Correa, Esteban G Burchard, Patrick T Ellinor, Stephen T McGarvey, Braxton D Mitchell, Ramachandran S Vasan, Susan Redline, Edwin Silverman, Scott T Weiss, Donna K Arnett, John Blangero, Eric Boerwinkle, Jiang He, Courtney Montgomery, D C Rao, Jerome I Rotter, Kent D Taylor, Jennifer A Brody, Yii-Der Ida Chen, Lisa de las Fuentes, Chii-Min Hwu, Stephen S Rich, Ani W Manichaikul, Josyf C Mychaleckyj, Nicholette D Palmer, Jennifer A Smith, Sharon L R Kardia, Patricia A Peyser, Lawrence F Bielak, Timothy D O'Connor, Leslie S Emery, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Population Genetics Working Group, Christian Gilissen, Wendy S W Wong, Peter V Kharchenko, and Shamil Sunyaev. Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science*, 373(6558):1030–1035, 2021. doi: 10.1126/science.aba7408.

Paul F Slade. Most recent common ancestor probability distributions in gene genealogies under selection. *Theoretical Population Biology*, 58(4):291–305, 2000a. doi: 10.1006/tpbi.2000.1488.

Paul F Slade. Simulation of selected genealogies. *Theoretical Population Biology*, 57(1):35–49, 2000b. doi: 10.1006/tpbi.1999.1438.

Montgomery Slatkin. Allele age and a test for selection on rare alleles. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1403):1663–1668, 2000. doi: 10.1098/rstb.2000.0729.

Montgomery Slatkin and Richard R Hudson. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129(2):555–562, 1991. doi: 10.1093/genetics/129.2.555.

Montgomery Slatkin and Bruce Rannala. Estimating allele age. *Annual Review of Genomics and Human Genetics*, 1(1):225–249, 2000. doi: 10.1146/annurev.genom.1.1.225.

Thomas Städler, Bernhard Haubold, Carlos Merino, Wolfgang Stephan, and Peter Pfaffelhuber. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, 182(1):205–216, 2009. doi: 10.1534/genetics.108.094904.

Matthew Stephens and Peter Donnelly. Ancestral inference in population genetics models with selection (with Discussion). *Australian & New Zealand Journal of Statistics*, 45(4):395–430, 2003. doi: 10.1111/1467-842X.00295.

Fumio Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2): 437–460, 1983. URL `https://www.genetics.org/content/105/2/437`.

Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123:585–595, 1989. doi: 10.1093/genetics/123.3.585.

Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, Achilleas N Pitsillides, Jonathon LeFaive, Seung-been Lee, Xiaowen Tian, Brian L Browning, Sayantan Das, Anne-Katrin Emde, Wayne E Clarke, Douglas P Loesch, Amol C Shetty, Thomas W Blackwell, Albert V Smith, Quenna Wong, Xiaoming Liu, Matthew P Conomos, Dean M Bobo, François Aguet, Christine Albert, Alvaro Alonso, Kristin G Ardlie, Dan E Arking, Stella Aslibekyan, Paul L Auer, John Barnard, R Graham Barr, Lucas Barwick, Lewis C Becker, Rebecca L Beer, Emelia J Benjamin, Lawrence F Bielak, John Blangero, Michael Boehnke, Donald W Bowden, Jennifer A Brody, Esteban G Burchard, Brian E Cade, James F Casella, Brandon Chalazan, Daniel I Chasman, Yii-Der Ida Chen, Michael H Cho, Seung Hoan Choi, Mina K Chung, Clary B Clish, Adolfo Correa, Joanne E Curran, Brian Custer, Dawood Darbar, Michelle Daya, Mariza de Andrade, Dawn L DeMeo, Susan K Dutcher, Patrick T Ellinor, Leslie S Emery, Celeste Eng, Diane Fatkin, Tasha Fingerlin, Lukas Forer, Myriam Fornage, Nora Franceschini, Christian Fuchsberger, Stephanie M Fullerton, Soren Germer, Mark T Gladwin, Daniel J Gottlieb, Xiuqing Guo, Michael E Hall, Jiang He, Nancy L Heard-Costa, Susan R Heckbert, Marguerite R Irvin, Jill M Johnsen, Andrew D Johnson, Robert Kaplan, Sharon L R Kardia, Tanika Kelly, Shannon Kelly, Eimear E Kenny, Douglas P Kiel, Robert Klemmer, Barbara A Konkle, Charles Kooperberg, Anna Köttgen, Leslie A Lange, Jessica Lasky-Su, Daniel Levy, Xihong Lin, Keng-Han Lin, Chunyu Liu, Ruth J F Loos, Lori Garman, Robert Gerszten, Steven A Lubitz, Kathryn L Lunetta, Angel C Y Mak, Ani Manichaikul, Alisa K Manning, Rasika A Mathias, David D McManus, Stephen T McGarvey, James B Meigs, Deborah A Meyers, Julie L Mikulla, Mollie A Minear, Braxton D Mitchell, Sanghamitra Mohanty, May E Montasser, Courtney Montgomery, Alanna C Morrison, Joanne M Murabito, Andrea Natale, Pradeep Natarajan, Sarah C Nelson, Kari E North, Jeffrey R O'Connell, Nicholette D Palmer, Nathan Pankratz, Gina M Peloso, Patricia A Peyser, Jacob Pleiness, Wendy S Post, Bruce M Psaty, D C Rao, Susan Redline, Alexander P Reiner, Dan Roden, Jerome I Rotter, Ingo Ruczinski, Chloé Sarnowski, Sebastian Schoenherr, David A Schwartz, Jeong-Sun Seo, Sudha Seshadri, Vivien A Sheehan, Wayne H Sheu, M Benjamin Shoemaker, Nicholas L Smith, Jennifer A Smith, Nona Sotoodehnia, Adrienne M Stilp, Weihong Tang, Kent D Taylor, Marilyn Telen, Timothy A Thornton, Russell P Tracy, David J Van Den Berg, Ramachandran S Vasan, Karine A Viaud-Martinez, Scott Vrieze, Daniel E Weeks, Bruce S Weir, Scott T Weiss, Lu-Chen Weng, Cristen J Willer, Yingze Zhang, Xutong Zhao, Donna K Arnett, Allison E Ashley-Koch, Kathleen C Barnes, Eric Boerwinkle, Stacey Gabriel, Richard Gibbs, Kenneth M Rice, Stephen S Rich, Edwin K Silverman, Pankaj Qasba, Weiniu Gan, Namiko Abe, Laura Almasy, Seth Ament, Peter Anderson, Pramod Anugu, Deborah Applebaum-Bowden, Tim Assimes, Dimitrios Avramopoulos, Emily Barron-Casella, Terri Beaty, Gerald Beck, Diane Becker, Amber Beitelshees, Takis Benos, Marcos Bezerra, Joshua Bis, Russell Bowler, Ulrich Broeckel, Jai Broome, Karen Bunting, Carlos Bustamante, Erin Buth, Jonathan Cardwell, Vincent Carey, Cara Carty, Richard Casaburi, Peter Castaldi, Mark Chaffin, Christy Chang, Yi-Cheng Chang, Sameer Chavan, Bo-Juen Chen, Wei-Min Chen, Lee-Ming Chuang, Ren-Hua Chung, Suzy Comhair, Elaine Cornell, Carolyn Crandall, James Crapo, Jeffrey Curtis, Coleen Damcott, Sean David, Colleen Davis, Lisa de las Fuentes, Michael DeBaun, Ranjan Deka, Scott Devine, Qing Duan, Ravi Duggirala, Jon Peter Durda, Charles Eaton, Lynette Ekunwe, Adel El Boueiz, Serpil Erzurum, Charles Farber, Matthew Flickinger, Chris Frazar, Mao Fu, Lucinda Fulton, Shanshan Gao, Yan Gao, Margery Gass, Bruce Gelb, Xiaoqi Priscilla Geng, Mark Geraci, Auyon Ghosh, Chris Gignoux, David Glahn, Da-Wei Gong, Harald Goring, Sharon Graw, Daniel Grine, C Charles Gu, Yue Guan, Namrata Gupta, Jeff Haessler, Nicola L Hawley, Ben Heavner, David Herrington, Craig Hersh, Bertha Hidalgo, James Hixson, Brian Hobbs, John Hokanson, Elliott Hong, Karin Hoth, Chao Agnes Hsiung, Yi-Jen Hung, Haley Huston, Chii Min Hwu, Rebecca Jackson, Deepti Jain, Min A Jhun, Craig Johnson, Rich Johnston, Kimberly Jones, Sekar Kathiresan, Alyna Khan, Wonji Kim, Greg Kinney, Holly Kramer, Christoph Lange, Ethan Lange, Leslie Lange, Cecelia Laurie, Meryl LeBoff, Jiwon Lee, Seunggeun Shawn Lee, Wen-Jane Lee, David Levine, Joshua Lewis, Xiaohui Li, Yun Li, Henry Lin, Honghuang Lin, Keng Han Lin, Simin Liu, Yongmei Liu, Yu Liu, James Luo, Michael Mahaney, and NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845):290–299, 2021. doi: 10.1038/s41586-021-03205-y.

Jonathan Terhorst and Yun S Song. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences*, 112 (25):7677–7682, 2015. doi: 10.1073/pnas.1503717112.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi: 10.1038/nature15393.

Raul Torres, Markus G Stetter, Ryan D Hernandez, and Jeffrey Ross-Ibarra. The temporal dynamics of background selection in nonequilibrium populations. *Genetics*, 214(4):1019–1030, 2020. doi: 10.1534/genetics.119.302892.

F Tricomi and A Erdélyi. The asymptotic expansion of a ratio of gamma functions. *Pacific Journal of Mathematics*, 1(1):133–142, 1951. doi: 10.2140/pjm.1951.1.133.

Claus Vogl and Florian Clemente. The allele-frequency spectrum in a decoupled Moran model with mutation, drift, and directional selection, assuming small mutation rates. *Theoretical Population Biology*, 81(3):197–209, 2012. doi: https://doi.org/10.1016/j.tpb.2012.01.001.

Claus Vogl, Lynette C Mikula, and Conrad J Burden. Maximum likelihood estimators for scaled mutation rates in an equilibrium mutation-drift model. *Theoretical Population Biology*, 134: 106–118, 2020. doi: 10.1016/j.tpb.2020.06.001.

G A Watterson. The sampling theory of selectively neutral alleles. *Advances in Applied Probability*, 6(3):463–488, 1974a. doi: 10.2307/1426228.

G A Watterson. Models for the logarithmic species abundance distributions. *Theoretical Population Biology*, 6(2):217–250, 1974b. doi: https://doi.org/10.1016/0040-5809(74)90025-2.

G A Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, 1975. doi: 10.1016/0040-5809(75)90020-9.

Carsten Wiuf. On the genealogy of a sample of neutral rare alleles. *Theoretical Population Biology*, 58(1):61–75, 2000. doi: https://doi.org/10.1006/tpbi.2000.1469.

Wolfram Research, Inc. Mathematica, Version 11.2, 2017. Champaign, IL.

Sewall Wright. Evolution in Mendelian populations. *Genetics*, 16(2):97–159, 1931. URL `https://www.genetics.org/content/16/2/97`.

Sewall Wright. The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences*, 24(7):253–259, 1938. doi: 10.1073/pnas.24.7.253.

Sewall Wright. Adaptation and selection. In G L Jepson, G G Simpson, and E Mayr, editors, *Genetics, Paleontology and Evolution*. Princeton Univ. Press, Princeton, 1949.