OXFORD GENETICS

GENETICS, 2023, 224(3), iyad049
https://doi.org/10.1093/genetics/iyad049
Advance Access Publication Date: 27 March 2023
Investigation

Recurrent mutation in the ancestry of a rare variant

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

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[2]Department of Mathematics, Indiana University, Bloomington, IN 47405, USA
[3]Center of Mathematical Sciences and Applications, Harvard University, Cambridge, MA 02138, USA
[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[5]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

Corresponding author: Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA. Email: wakeley@fas.harvard.edu
†These authors contributed equally to this work.

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 distributed like the number of alleles in 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.

Keywords: recurrent mutation, Ewens sampling formula, coalescent theory, human SNPs

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 *et al.* 2022) 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.

Received: January 30, 2023. **Accepted:** March 8, 2023

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) assumed 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 61 000 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. In short, the distributions of latent mutations and counts of rare variants depend on the expected total length of the gene genealogy of the sample, the expected lengths of branches with few descendants in the sample, and of course the mutation rate. 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 57 K individuals, pre-sorted into 97 bins based on estimates of mutation rate by the method of Seplyarskiy *et al.* (2022).

We develop and present these results in the next three sections. In “Theory for constant-size large populations,” 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 “Theory for nonconstant populations,” we extend the results to populations which have changed in size, using the Poisson-sampling models of Watterson (1974b) and Arratia *et al.* (1992). In “Theoretical example and data application,” 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.

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 numbers are distributed like the numbers of alleles in the Ewens sampling formula.

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 through $K-1$, and necessarily $n_K = n - \sum_{i<K} n_i$ copies of allele K, with probability

$$p(n_1, \ldots, n_{K-1}; n) \equiv \mathbb{P}[\mathcal{N}_1 = n_1, \ldots, \mathcal{N}_{K-1} = n_{K-1}; n]$$

$$= \binom{n}{n_1 \cdots n_K} \mathbb{E}[X_1^{n_1} \cdots X_{K-1}^{n_{K-1}}] \tag{2}$$

$$= \binom{n}{n_1 \cdots n_K} (\theta^{(n)})^{-1} \prod_{i=1}^{K} (\theta\pi_i)^{(n_i)} \tag{3}$$

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 (3) 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

$$\phi(x) = \frac{\Gamma(\theta)}{\Gamma(\theta\pi_1)\Gamma(\theta\pi_2)} x^{\theta\pi_1 - 1}(1-x)^{\theta\pi_2 - 1} \tag{4}$$

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

$$p(n_1; n) = \binom{n}{n_1} \frac{(\theta\pi_1)^{(n_1)}(\theta\pi_2)^{(n-n_1)}}{\theta^{(n)}} \tag{5}$$

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 (5) 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, 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 (The 1000 Genomes Project Consortium 2015) 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, 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.

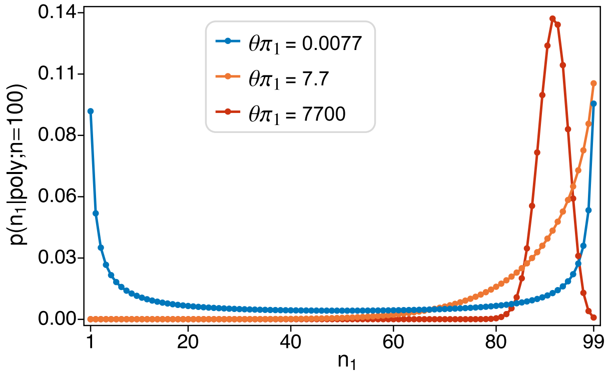

Fig. 1. Sample frequency distribution $p(n_1; n)$ for $n = 100$, with $\pi_1 = 10\pi_2$ and three values of θ. The smallest θ was chosen so that $\theta\pi_2 = 1/1300 \sim 0.00077$, i.e. the human average value. Probabilities are normalized to sum to one, i.e. conditioned on the sample being polymorphic ($1 \leq n_1 \leq 99$).

Relationship to infinite-sites frequency spectra

We use θ for the per-site mutation parameter. In a collection of L total sites at which (5) 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 by the product $Lp(n_1; n)$. Note, these expected numbers of sites do not depend on the rate of recombination, whereas the variances among sites and covariances between sites do (Kaplan and Hudson 1985).

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 (4) we may write

$$\phi_{ISM}(x) \propto \frac{\theta\pi_1}{x} \tag{6}$$

for a single site (θ small) approximately under the standard infinite-sites mutation model. For comparison with (5), we have

$$p_{ISM}(n_1; n) \propto \frac{\theta\pi_1}{n_1} \tag{7}$$

for the approximate single-site probability that there are n_1 type-1 alleles in a sample of size n. Equation (7) 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 (4) and (5) with θ small we have

$$\phi(x) = \pi_2 \frac{\theta\pi_1}{x} + \pi_1 \frac{\theta\pi_2}{1-x} + O(\theta^2) \tag{8}$$

and

$$p(n_1; n) = \pi_2 \frac{\theta\pi_1}{n_1} + \pi_1 \frac{\theta\pi_2}{n_2} + O(\theta^2) \tag{9}$$

for $n_1 \in \{1, \ldots, n-1\}$. The diffusion result (4) 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 (8) 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 (8) represent a mixture of two infinite-sites models like (6) with the constants of proportionality specified. Equation (9) has an identical interpretation, as a mixture of two infinite-sites site-frequency spectra. These are the key principles of the boundary mutation model (Vogl and Clemente 2012; Vogl *et al.* 2020).

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 (8) when θ is small; see equation (21) in Burden and Tang (2017). See also Schrempf and Hobolth (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 (9) 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 (9).

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.

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 $p_o(n_1, \ldots, n_K)$ in place of $p_o(n_1, \ldots, n_{K-1}; n)$ to facilitate the calculations, we have

$$p_o(n_1, \ldots, n_K) = (\theta^{(n)})^{-1} \prod_{i=1}^{K} (\theta\pi_i)^{(n_i)} \qquad (10)$$

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

Equation (10) is suggestive, as are (3) and (5), 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

$$(\theta\pi_i)^{(n_i)} = \sum_{k_i=1}^{n_i} \left|S_{n_i}^{(k_i)}\right| (\theta\pi_i)^{k_i}, \qquad (11)$$

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 $k_i \in \{1, \ldots, n_i\}$ most recent ones which produced the observed alleles.

That is, based on (10) and (11), we suppose that the joint probability of the sample counts $n_1, \ldots, n_K$ and their numbers of latent mutations $k_1, \ldots, k_K$ is given by

$$p_o(k_1, \ldots, k_K, n_1, \ldots, n_K) = (\theta^{(n)})^{-1} \prod_{i=1}^{K} \left|S_{n_i}^{(k_i)}\right| (\theta\pi_i)^{k_i}, \qquad (12)$$

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

$$p(k_1, \ldots, k_K | n_1, \ldots, n_K) = \prod_{i=1}^{K} \frac{\left|S_{n_i}^{(k_i)}\right| (\theta\pi_i)^{k_i}}{(\theta\pi_i)^{(n_i)}} \qquad (13)$$

which applies to both ordered and unordered samples.

We show that (13) is true using the ancestral-process approach of Griffiths and Tavaré (1994a, 1994b). If sampling probabilities like (3) or (10) are known, this approach can be used to describe the conditional ancestral process of a sample given its allelic types (Slade 2000a, 2000b; Fearnhead 2001, 2002; Stephens and Donnelly 2003; Baake and Bialowons 2008). Following our analysis of (13) for arbitrary $(n_1, \ldots, n_K)$, we describe a large-n approximation in which allele K is the overwhelmingly common type and 1 through $K-1$ are the rare variants.

The conditional ancestral process has the same total rate of mutation and coalescence as the unconditional process, $n(\theta + n - 1)/2$. Lineages which must be of type i in the sample experience type-i mutations at rate $n_i\theta\pi_i/2$ and type-i coalescent events at rate $n_i(n_i - 1)/2$, but with additional weights proportional to the probability of $(n_1, \ldots, n_K)$ given each event. All other events have rates equal to zero because the sample could not be $(n_1, \ldots, n_K)$ if they occurred. To obtain (13), we follow ancestral lineages only back to the first mutation event they experience. The probability of a type-i mutation event is

$$\frac{n_i\theta\pi_i p_o(\ldots, n_i - 1, \ldots)}{n(\theta + n - 1)p_o(n_1, \ldots, n_K)} = \frac{n_i}{n}\frac{\theta\pi_i}{\theta\pi_i + n_i - 1}, \qquad (14a)$$

and the probability of a type-i coalescent event is

$$\frac{n_i(n_i - 1)p_o(\ldots, n_i - 1, \ldots)}{n(\theta + n - 1)p_o(n_1, \ldots, n_K)} = \frac{n_i}{n}\frac{n_i - 1}{\theta\pi_i + n_i - 1}, \qquad (14b)$$

where we have used (10) to obtain the results on the right. Whether mutation or coalescence occurs, the number of type i lineages decreases by one: $n_i \rightarrow n_i - 1$. This ancestral process continues until there are no un-mutated ancestral lineages, that is until $n_i = 0$ for all $i \in \{1, \ldots, K\}$.

To this we add a mutation counting process which starts with $k_i = 0$ for all $i \in \{1, \ldots, K\}$ then has $k_i \rightarrow k_i + 1$ whenever a mutation occurs on a type-i ancestral lineage. Equations (14a) and (14b) show that each event in the ancestral process includes two sub-events: a choice of the allelic type involved then a choice between mutation and coalescence. Depending on $(n_1, \ldots, n_K)$, the $n = \sum_i n_i$ choices of allelic type will result in a random ordering of events among types. But for every ordering, the series of choices between mutation and coalescence within allelic type i depends only on n_i (and $\theta\pi_i$) and is independent of what happens in the ancestry of allele $j \neq i$. The number of mutations of type i is the sum of n_i Bernoulli random variables with success probabilities $\theta\pi_i/(\theta\pi_i + j - 1)$ for j from n_i down to 1. The number of latent mutations counted in this way will be distributed like the number of alleles in the Ewens sampling formula—see Arratia *et al.* (1992) and Arratia and Tavaré (1992)—with mutation parameter $\theta\pi_i$ for allele i, and these counts will be independent among alleles as in (13).

We use this conditional ancestral process below but here note its close relationship to models of lines of descent (Griffiths 1980; Watterson 1984). In particular, (13) is included in equation (3.3) and Theorem 4 of Donnelly (1986), who extended Watterson's lines-of-descent model to the case of K-allele, parent-independent mutation. See also Donnelly and Tavaré (1987). Equation (3.3) in Donnelly (1986) in fact shows that if we were to keep track of the numbers of descendants of each latent mutation, the full Ewens sampling formula would give their distribution in the sample.

Before describing a large-n approximation for rare variants, we also note that latent mutations reckoned as in (13) include what Donnelly (1986) called 'spurious mutations to one's own type' and Baake and Bialowons (2008) called 'empty mutations'. These are a modeling artifact not only of parent-independent mutation models but of general mutation models as they 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 and should not be counted as mutations. To deal with them, we must keep track of the ancestral types of lineages when they experience mutations. We can do using the identity

$$p_o(\ldots, n_i - 1, \ldots) = \sum_{j=1}^{K} p_o(\ldots, n_i - 1, \ldots, n_j + 1, \ldots) \qquad (15)$$

which decomposes our previously generic type-i mutations according to their ancestral types $j \in \{1, \ldots, K\}$. A mutation is empty when $j = i$.

In our large-n approximation, we take K to be the overwhelmingly common allelic type in the sample and 1 through $K-1$ to be the rare variants. Our goal is to model latent mutations in the ancestry of the rare variants, so we use (15) only for $i \in \{1, \ldots, K-1\}$. For the common allele K, we instead lump (14a) and (14b) together and record both mutation and coalescence as $n_K \to n_K - 1$. Making these changes to (14a) and (14b), and again using (10) to simplify ratios of sampling probabilities, the conditional ancestral process for a sample with state $(n_1, \ldots, n_K)$ jumps to state $(\ldots, n_i - 1, \ldots, n_j + 1, \ldots)$ for $i, j \neq K$ with probability

$$\frac{n_i}{n} \frac{\theta\pi_i(\theta\pi_j + n_j - \delta_{ij})}{(\theta + n - 1)(\theta\pi_i + n_i - 1)}, \qquad (16a)$$

to state $(\ldots, n_i - 1, \ldots, n_K + 1)$ for $i \neq K$ with probability

$$\frac{n_i}{n} \frac{\theta\pi_i(\theta\pi_K + n_K)}{(\theta + n - 1)(\theta\pi_i + n_i - 1)}, \qquad (16b)$$

to state $(\ldots, n_i - 1, \ldots)$ for $i \neq K$ with probability

$$\frac{n_i}{n} \frac{n_i - 1}{\theta\pi_i + n_i - 1}, \qquad (16c)$$

and to state $(\ldots, n_K - 1)$ with probability

$$\frac{n_K}{n} \qquad (16d)$$

where we have used Kronecker's delta to accommodate empty mutations, $i = j$ in (16a). Equation (16a) includes both empty and nonempty mutations, but only ones where the ancestral type is also rare. Nonempty mutations where the ancestral type is the common type K are in (16b). This classification of mutations by ancestral type does not change the probabilities of coalescence, so (16c) only differs from (14b) by the absence of type-K coalescent events which are now in (16d).

If n_K is large compared to n_1 through n_{K-1}, then $n = \sum_i n_i \approx n_K$. The probabilities in (16a) will be $O(1/n_K^2)$, those in (16b) and (16c) will be $O(1/n_K)$, and the one in (16d) will be $O(1)$. Empty mutations and other mutations with rare-variant ancestors will become negligible as n_K grows for fixed n_1 through n_{K-1}. Keeping only terms of $O(1/n_K)$ and larger gives an approximate, large-n ancestral process with total rate $n(\theta + n - 1)/2 \approx n_K^2/2$ and jumps, for $i \in \{1, \ldots, K-1\}$, from state $(n_1, \ldots, n_K)$ to state $(\ldots, n_i - 1, \ldots, n_K + 1)$ with probability

$$\frac{n_i}{n_K} \frac{\theta\pi_i}{\theta\pi_i + n_i - 1}, \qquad (17a)$$

to state $(\ldots, n_i - 1, \ldots)$ with probability

$$\frac{n_i}{n_K} \frac{n_i - 1}{\theta\pi_i + n_i - 1}, \qquad (17b)$$

and to state $(\ldots, n_K - 1)$ with probability

$$1 - \frac{\sum_{i=1}^{K-1} n_i}{n_K}. \qquad (17c)$$

This process is dominated by (17c), 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 (17a) or a coalescent event as in (17b). This process for the rare variants $i \in \{1, \ldots, K-1\}$ has the same form as that found for all variants and all mutations in (14a) and (14b). Then by the same logic as before, the number of (now nonempty) latent mutations in the ancestry of the rare variants will be distributed like the number of alleles in the Ewens sampling formula, independently and with mutation parameter $\theta\pi_i$ for allele $i \in \{1, \ldots, K-1\}$. In addition if we were to keep track of the counts of each mutation's descendants among the n_i copies of rare variant i in the sample, then because every pair of type-i lineages is equally likely to be the one which coalesces when a type-i coalescent event occurs, the distribution of these counts should be given by the full Ewen's sampling formula (Ewens 1972; Kingman 1982; Donnelly 1986; Arratia and Tavaré 1992; Arratia *et al.* 1992, 2016).

The events involving the common allele in (17c) 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 "Time-dependent conditional ancestral process," 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

$$\mathbb{E}[\mathcal{T}_1] \approx \begin{cases} \frac{2\log(n_2)}{n_2} & \text{if } n_1 = 1 \\ \frac{2}{n_2(n_1-1)} & \text{if } n_1 > 1 \end{cases} \tag{18}$$

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

$$\mathbb{E}[\mathcal{N}_2(\mathcal{T}_i)] \approx n_2 \frac{n_1 - i + 1}{n_1 + 1} \tag{19}$$

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 ($n_2 \sim 86,000$) and >400 left for the gnomAD data ($n_2 \sim 30,000$). In section "Application to human SNP data," we consider site-frequency counts up to 40 for synonymous exonic sites in gnomAD with many fewer SNPs but a larger sample size ($n_2 \sim 114,000$) 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.

In sum, 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 these latent mutations in the ancestry of allele $i \in \{1, \ldots, K-1\}$, then from the rates of mutation and coalescence in (17a) and (17b) we have

$$p(k_1, \ldots, k_{K-1} | n_1, \ldots, n_{K-1}; n \text{ large}) \approx \prod_{i=1}^{K-1} \frac{\left|S_{n_i}^{(k_i)}\right| (\theta\pi_i)^{k_i}}{(\theta\pi_i)^{(n_i)}}. \tag{20}$$

Latent mutations of different rare variants are independent and distributed like the numbers of alleles in the Ewens sampling formula, each with its own mutation parameter.

Latent mutations and sample counts of rare alleles

Our goal in this section is to understand how predictions about the counts of rare variants, and hence about their site-frequency spectra, depend on the number of latent mutations and the mutation rate. In anticipation of "Application to human SNP data," we focus on the marginal count of just one rare variant, which we arbitrarily call allele 1. From (20) we have

$$p(k_1 | n_1; n \text{ large}) \approx \frac{\left|S_{n_1}^{(k_1)}\right| (\theta\pi_1)^{k_1}}{(\theta\pi_1)^{(n_1)}}, \quad k_1 \in \{1, \ldots, n_1\} \tag{21}$$

which we note holds for any K. Here we let $K = 2$ for simplicity.

To understand how the mutation rate influences the count of a rare variant, we apply the result for ratios of gamma functions with a common large parameter, 6.1.47 in Abramowitz and Stegun (1964) or equation (1) in Tricomi and Erdélyi (1951), to the terms involving n in (5) to obtain

$$p(n_1; n) = \frac{(\theta\pi_1)^{(n_1)}}{n_1!} e^{-\theta\pi_1 \log(n)} \frac{\Gamma(\theta)}{\Gamma(\theta\pi_2)} \left[1 + O\left(\frac{1}{n}\right)\right], \tag{22}$$

in which we have used $n^{-\theta\pi_1} = e^{-\theta\pi_1 \log(n)}$ to make a connection with the underlying coalescent tree or gene genealogy. Specifically, $\theta\pi_1 \sum_{i=1}^{n-1} 1/i$ is the expected number of type-1 mutations on the gene genealogy of a sample of size n, and for large n this is approximately equal to $\theta\pi_1(\log(n) + \gamma)$ where $\gamma = 0.5772\ldots$ is Euler's constant. In "Theory for nonconstant populations" we explore this connection in detail and explain the additional constants of proportionality in (22) after finding an analogous result for the general coalescent trees of Griffiths and Tavaré (1998).

Site-frequency spectra are typically defined as the proportion of segregating sites in each possible count in the sample (Braverman *et al.* 1995) or equivalently as the probability that a single mutation is in each possible count given that it is polymorphic in the sample (Griffiths and Tavaré 1998; Nielsen 2000). So, to understand how n_1 depends on $\theta\pi_1$, we may ignore the constants of proportionality in (22) and focus on

$$p(n_1; n \text{ large}) \propto \frac{(\theta\pi_1)^{(n_1)}}{n_1!}. \tag{23}$$

Then using (23) together with (21), we have

$$p(n_1 | k_1; n \text{ large}) \propto \frac{\left|S_{n_1}^{(k_1)}\right|}{n_1!} \tag{24}$$

for the dependence of the rare-variant count, n_1, on the number of latent mutations, k_1, relevant to the site-frequency spectrum. Figure 2 shows site-frequency spectra computed using (23) and (24), and conditioning on the event that $n_1 \in \{1, 2, \ldots, 40\}$.

Figure 2a shows the dependence on the number of latent mutations. When all copies descend from a single mutation ($k_1 = 1$), the usual predictions from the infinite-sites model hold. Thus if we put $|S_{n_1}^{(1)}| = (n_1 - 1)!$ in (24), then consistent with (7) 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 (21) for larger numbers of latent mutations. But conditional on k_1, the site-frequency counts for a rare variant do not depend on θ, at least to leading order in the sample size n. 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 shown in Fig. 2a for $k_1 = 2$ to $k_1 = 5$. 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 (21) and (23) the joint distribution of n_1 and k_1 obeys

$$p(n_1, k_1; n \text{ large}) \propto \frac{\left|S_{n_1}^{(k_1)}\right| (\theta\pi_1)^{k_1}}{n_1!} \tag{25}$$

which can be compared to the results of Jenkins and Song (2011). With fixed n_1 and large n in our model, all mutations in the ancestry of the rare variant will be non-nested mutations; note this also

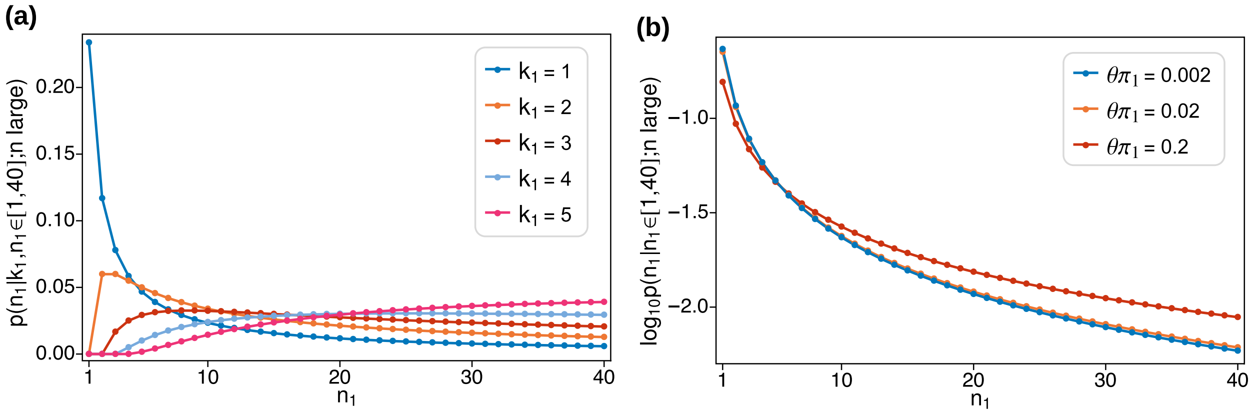

Fig. 2. a) Shows the probability of observing n_1 copies of allele 1 in a large sample given these are produced by k_1 mutations. b) Shows the $\log_{10}$-probability of observing n_1 copies of allele 1 in a large sample for three different values of $\theta\pi_1$. In both panels, probabilities are normalized to sum to one, that is conditioned on the event that $n_1 \in \{1, 2, \ldots, 40\}$.

follows from (18) in Jenkins and Song (2011). Adapting the notation of Jenkins and Song (2011) in which $E^{(1,1)}_{2\mathcal{N},\mathcal{N}}$ is the event that the n_1 copies of allele 1 are due to two non-nested mutations, both from allele $K = 2$ to allele 1, their (21) becomes

$$p(n_1, n_2, E^{(1,1)}_{2\mathcal{N},\mathcal{N}}) \approx \theta^2\pi_1^2 \frac{\left|S^{(2)}_{n_1}\right|}{n_1!}$$

for large $n \sim n_2$ (and small θ), which is identical to (25) 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 Fig. 2a when n is large. This is evident in Figure 3 of Jenkins and Song (2011) for the quantity $E_{2\mathcal{N}N}$. 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 Fig. 2a shows a dramatic effect of k_1 on the site-frequency counts, Fig. 2b suggests that large values of k_1 are unlikely. This is evident from (21) and (25) in that each additional mutation results in an additional factor of $\theta\pi_1$. Note that the smallest value of $\theta\pi_1$ in Fig. 2b is already more than twice the human average. From (23), we have

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

which is consistent with (9) 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 Fig. 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 Fig. 2b) the site-frequency spectrum under recurrent mutation is noticeably 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.

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. Here gene genealogies are the general coalescent trees of Griffiths and Tavaré (1998), which have the same branching structure of standard coalescent trees but may have different distributions of coalescence times.

Equation (21) 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, …, 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 (1974b) 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 (21) 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,

$$\mathbb{E}[\tau_i] = \sum_{k=2}^{n} k\mathbb{E}[T_k] \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \qquad (26)$$

(Fu 1995; Griffiths and Tavaré 1998).

Watterson (1974b) 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 (1974b) proved that the numbers and counts of mutations follow the Ewens sampling formula when conditioned on their total size, which for Watterson (1974b) was equivalent to the population size. Models 2 and 3 are the Moran model and the Wright-Fisher model (Fisher 1930b; Wright 1931; Moran 1958, 1962) and Watterson (1974b) 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 1974a; 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 (1974b) to Model 1. Whereas Watterson (1974b) 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 (1974b) 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

$$\begin{aligned} P(a_1, a_2, \ldots) &= \prod_{i\geq 1} \frac{(\theta\bar{\tau}_i/2)^{a_i}}{a_i!} e^{-\theta\bar{\tau}_i/2} \\ &= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \prod_{i\geq 1} \frac{(\theta\bar{\tau}_i/2)^{a_i}}{a_i!} \end{aligned} \tag{27}$$

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 (27), 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 "Low-count branches of general coalescent trees" 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 finite, 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 (1974b), we consider the probability generating function of n and k, which in the present case simplifies to

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

For the details of this derivation, see (A29) in the Appendix. The coefficient of x^n (and y^k) can be found using

$$\left(\sum_i x^i \bar{\tau}_i\right)^k = \sum_{n\geq k} x^n \sum_{(i_1,\ldots,i_{k-1})} \bar{\tau}_{i_1}\bar{\tau}_{i_2}\cdots\bar{\tau}_{i_k} \tag{28}$$

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 $\boldsymbol{\tau}$ to show the new dependence on the vector of expected times $(\bar{\tau}_1, \ldots, \bar{\tau}_{n-1})$, we have

$$p(n_1, k_1; n \text{ large}, \boldsymbol{\tau}) \approx \frac{\left(\frac{\theta\pi_1}{2}\right)^{k_1} \sum_{(i_1,\ldots,i_{k_1-1})} \prod_{m=1}^{k_1} \bar{\tau}_{i_m}}{k_1!} e^{-\frac{\theta\pi_1}{2}\sum_{i=1}^{n-1}\bar{\tau}_i} \tag{29}$$

which is nonzero 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 (28). 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 (29) as follows. The number of type-1 mutations is Poisson distributed

$$p(k_1; n \text{ large}, \boldsymbol{\tau}) \approx \frac{\left(\frac{\theta\pi_1}{2}\sum_{i=1}^{n-1}\bar{\tau}_i\right)^{k_1}}{k_1!} e^{-\frac{\theta\pi_1}{2}\sum_{i=1}^{n-1}\bar{\tau}_i}, \tag{30}$$

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

$$p(n_1|k_1; n \text{ large}, \boldsymbol{\tau}) \approx \sum_{(i_1,\ldots,i_{k_1-1})} \prod_{m=1}^{k_1} \frac{\bar{\tau}_{i_m}}{\sum_{i=1}^{n-1}\bar{\tau}_i}, \tag{31}$$

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 (29) appropriately, over $k_1 \in (0, \ldots, n_1)$. Then

$$p(k_1|n_1; n \text{ large}, \tau) \approx \frac{p(n_1, k_1; n \text{ large}, \tau)}{p(n_1; n \text{ large}, \tau)} \tag{32}$$

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 (31) and (29) 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 each site. Equation (30) 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 Fig. 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 (31) 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 (31),

$$p(n_1|0; n \text{ large}, \tau) \approx \begin{cases} 1 & \text{if } n_1 = 0 \\ 0 & \text{if } n_1 \geq 1 \end{cases} \tag{33}$$

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

$$p(n_1|2; n \text{ large}, \tau) \approx \frac{\sum_{i=1}^{n_1-1}\bar{\tau}_i\bar{\tau}_{n_1-i}}{\left(\sum_{i=1}^{n-1}\bar{\tau}_i\right)^2} \tag{35}$$

Equation (33) says simply that if there are no type-1 mutations on the gene genealogy then no copies of allele-1 will be observed. Equation (34) is the familiar result for the site-frequency spectrum, that it is given by the proportion of branches in the tree that have n_1 descendants. Equation (35) 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 find new approximations

$$p(n_1; n \text{ large}, \bar{\tau}_i = 2/i) \approx \frac{(\theta\pi_1)^{(n_1)}}{n_1!}e^{-\theta\pi_1\sum_{i=1}^{n-1}1/i} \tag{36}$$

$$p(n_1|k_1; n \text{ large}, \bar{\tau}_i = 2/i) \approx \frac{\left|S_{n_1}^{(k_1)}\right|k_1!}{n_1!}\left(\sum_{i=1}^{n-1}\frac{1}{i}\right)^{-k_1} \tag{37}$$

$$p(n_1, k_1; n \text{ large}, \bar{\tau}_i = 2/i) \approx \frac{\left|S_{n_1}^{(k_1)}\right|(\theta\pi_1)^{k_1}}{n_1!}e^{-\theta\pi_1\sum_{i=1}^{n-1}1/i} \tag{38}$$

corresponding to (23), (24) and (25), respectively, in which the condition $\bar{\tau}_i = 2/i$ should be taken to hold for all $i \in \{2, \ldots, n\}$. Figure 2 is unchanged if (36) and (37) are used instead of (23) and (24). Also, the conditional probability of k_1 given n_1 from (36) and (38) is identical to (21).

Relation to *K*-alleles diffusion results

From a gene-genealogical point of view, (36) is the probability of seeing n_1 total copies of a rare variant when a random number of type-1 mutations occurs on the low-count branches of a standard neutral coalescent tree. However, the type of the common variant and the ancestral states of these mutations are not specified in the independent-Poissons model. Of course these should be allele K, as in "A conditional ancestral process for rare variants," but (36) does not include this event. In contrast, the sampling probabilities (3) and (5) from the equilibrium diffusion model specify the types of the entire sample. Implicitly, they average over the ancestral states of the sample. Here we focus on $K = 2$ and show how (36) is related to (5) when n is large, in particular to the leading order term in the expansion (22).

The type of the common ancestor of the entire sample, at the root of the coalescent tree, is allele 2 with probability π_2. If this were the case, allele 2 would be the ancestral source of the low-count type-1 mutations. But if θ is not very small, it is possible for allele 2 to be the ancestral source of these mutations even if the common ancestor is type 1. To illustrate, dividing either (5) or (22) by (36) and letting $n \to \infty$ gives

$$\frac{e^{\theta\pi_1\gamma}\Gamma(\theta)}{\Gamma(\theta\pi_2)} = \pi_2 + O(\theta^2). \tag{39}$$

Indeed when θ is small, (22) is close to (36) times π_2. But the error of this, even as n tends to infinity, may be appreciable for larger values of θ. The additional probability of order θ^2 in (39) is consistent with the possibility that the root of the coalescent tree is type 1 and there are two type-2 mutations, one on each of the two branches descending from the root.

A better guarantee that allele 2 is the ancestral source of low-count mutations would be to specify it not as type of the single most recent common ancestor but rather as the type of the pair of ancestors at the first time in the past when there were two ancestral lineages. Equation (5), with sample size equal to two, gives the relevant probability. This accounts for both possible states at the root of the tree as well as for mutation during the deepest coalescent interval, T_2 in (26). Then the independent-Poissons model could be applied to the remainder of the tree, i.e. to coalescent intervals T_3 through T_n.

Because latent mutations of rare variants tend to be very recent, cf. (18) and (19), we may extend this logic to the first time in the past when there were r ancestral lines of the sample, for an arbitrary $r \geq 1$. The probability that these are all of type 2 is given by the diffusion result (5) with sample size r. The probability of seeing n_1 copies of the rare variant is given by an appropriately adjusted independent-Poissons model, covering coalescent intervals T_{r+1} through T_n. By summing (26) only over $j \in \{r+1, \ldots, n\}$ it can be shown that the total length of branches with i descendants in this more recent part of the gene genealogy differs only by $2(1-r)/n + O(1/n^2)$ from the full result $\bar{\tau}_i = 2/i$. The product of these two probabilities is

$$\frac{\Gamma(\theta)\Gamma(\theta\pi_2 + r)}{\Gamma(\theta\pi_2)\Gamma(\theta + r)}\frac{(\theta\pi_1)_{n_1}}{n_1!}e^{-\theta\pi_1\sum_{i=r}^{n-1}1/i} \tag{40}$$

which can be compared to the leading order term in (22).

As expected from (39), if $r = 1$ (40) reduces to (36) times π_2. Now dividing (5) or (22) by (40) and letting $n \to \infty$ gives

$$\frac{\Gamma(\theta\pi_2 + r)}{\Gamma(\theta + r)}e^{-\theta\pi_1\left(\sum_{i=1}^{r-1}1/i-\gamma\right)} \tag{41}$$

as a measure of how well this augmented independent-Poissions model approximates the equilibrium diffusion result, depending on r and θ. Expanding (41) around $\theta = 0$, because we do not in fact expect the per-site mutation parameter to be large, gives

$$1 + \frac{(2-\pi_1)\pi_1}{2}\left(\frac{\pi^2}{6} - \sum_{j=1}^{r-1}\frac{1}{j^2}\right)\theta^2 + O(\theta^3) \tag{42}$$

where the π in π^2 is the usual constant (not our π_i). The parenthetic term in (42) tends to zero quickly as r increases. It is equal to the trigamma function $\psi^{(1)}(r)$ for $r \in \{1, 2, 3, \ldots\}$; see 6.4.2 and 6.4.3 in Abramowitz and Stegun (1964). Even just taking $r = 2$ instead of $r = 1$ cuts the error by about 60%.

Similar conclusions may be drawn from the large-r expansion of (41), which gives $1 + (2 - \pi_1)\pi_1\theta^2/(2r) + O(1/r^2)$. Again θ^2 is the largest-order effect of mutation. The event that a pair of mutations occurs on the two lineages descending from the root of the coalescent tree is non-negligible in the constant-size population model, even as $n \to \infty$ and even for the entire population, because ancient coalescence times tend to be long. But the chance of this event will be small for most eukaryote species as θ ranges from about 10^{-4} to 10^{-1} with typical values around 10^{-2} (Leffler *et al.* 2012). Based on our estimates in the next section, even the fastest-mutating sites in the human genome have $\theta \approx 0.02$. Note that this event will even less likely in growing populations, because in this case the deepest coalescence times will be relatively short, but could be an important phenomenon for populations which were much larger in the past.

Theoretical example and data application

Here we illustrate the theoretical and empirical use of (30) and (31). 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 (31). 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 (30) and (31).

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 $\beta = N_0 r$,

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

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 of 30 years and $r = 0.0064$, would have $\beta = 3.2 \times 10^5$. 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 (26) 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

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

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 $x = k/n$,

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

we approximate the expected coalescent interval as

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

Note that while (48) is a large-n approximation, it allows that β might be of the same order of magnitude as n. Applying the same approximation to the combinatorial coefficient in (26) 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 (26) with the integral

$$\mathbb{E}[\tau_i] \approx \int_0^1 xn\frac{2}{x(2\beta(1-x) + xn)}\frac{x}{1-x}(1-x)^i\,dx,$$

$$= \frac{n}{\beta}\int_0^1\left[1 - \left(1 - \frac{n}{2\beta}\right)x\right]^{-1} x(1-x)^{i-1}dy \tag{50}$$

$$= \frac{n}{\beta i(i+1)}\ {}_2F_1\left(1, 2;\, i+2;\, 1 - \frac{n}{2\beta}\right) \tag{51}$$

which can be evaluated efficiently either as (51), in terms of the hypergeometric function, or as the integral (50). Slatkin and Hudson (1991) and others have observed that gene genealogies under very fast exponential growth are close to star trees. Using either (50) or (51) we have

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

as β/n increases. From the $\log(2\beta/n)$ term in (52), we confirm the star-tree prediction that under extreme growth essentially all variants will be singletons.

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; Durrett 2013, 2015; Kessler and Levine 2013; Ohtsuki and Innan

2017; Cheek and Antal 2018; Gunnarsson et al. 2021; Poon et al. 2021). 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 3 shows the same quantities as Fig. 2 but for the pure exponential growth model with $n = 10^5$ and $\beta/n = 3$. The value $\beta/n = 3$ was chosen to roughly reproduce the ratio of singletons to doubletons observed for low-rate sites in the gnomAD data in section "Application to human SNP data." Figure 3a is directly comparable to Fig. 2a, the only difference being whether $\mathbb{E}[\tau_i] = 2/i$ or comes from (51). As Fig. 3a 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 Fig. 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.

To make Fig. 3b comparable to Fig. 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 Fig. 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 "Application to human SNP data." We then computed $p(n_1; n \text{ large}, \tau)$ by averaging (31) over the distribution (30). Similar to Fig. 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 Fig. 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 Fig. 3b 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 (Fig. 3a) than under constant size (Fig. 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 4 depicts this for a series of increasing counts n_1, from 1 to 16. Figure 4a shows the results for constant size, Fig. 4b 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.

Application to human SNP data

We also used (30) and (31) to account for latent mutations in the ancestry of rare 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 the data at low-rate sites, then using our new analytical result (31) 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 mutation rate by the Roulette method of Seplyarskiy et al. (2022) which incorporates information from the six flanking bases on either side of a SNP, strand asymmetry, expression level, methylation and promoter status. We did not use this information but simply assumed that variants within a bin all have the same mutation rate.

The data consist of variant counts for synonymous mutations in the exomes of about 57 K non-Finnish Europeans. Thus $n \sim 114$ K 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. The gnomAD only provides n for polymorphic sites, so we imputed n for monomorphic sites using the nearest value at a polymorphic site within 100 bp 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 5a shows the total numbers of sites and the numbers of monomorphic sites in each bin. The great majority of sites are in bins 1 through roughly 20. These have low mutation rates, as indicated by their nearly equal numbers of total sites and monomorphic sites. The widening gap between the total number of sites and the number of monomorphic sites reflects the fact that higher-number bins have larger mutation rates.

For each bin, the data are the numbers of sites where a variant is observed in each possible count in the sample. As in "Latent mutations and sample counts of rare alleles," these are marginal with respect to other possible variants at the site. Sites with two (resp. three) rare variants appear twice (resp. three times) in the data, once for each rare variant. These will likely be in different bins given the fine substructure of mutation rate variation (Seplyarskiy et al. 2021, 2022). Although bins contain mixtures of different sequence contexts and different nucleotide substitutions, for our purposes sites within a bin are all of the same type because they all have the same mutation rate.

Let S_i be the number of sites in a given bin where i copies of the variant are observed in the sample. If a bin contains L total sites, then with reference to the notation in (2) we may write

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

Thus we use a simplified notation here, with i in place of n_1 to avoid the additional subscript when we apply the results of the previous sections. In addition we use "mutrate" to refer to the estimate of the expected number of latent mutations per site for a given bin, i.e. $(\theta\pi_1/2)\sum_i \bar{\tau}_i$ for sites in that bin, as this is the rate parameter in the Poisson distribution (30).

We used (30) and the proportion of monomorphic sites, S_0/L, to estimate this "mutrate" for each bin, specifically as $-\log(S_0/L)$. Figure 5b plots these estimates across bins, on a log scale. They range from 0.0097 for bin 1 to 2.23 for bin 97, with a mean of 0.083, taking the proportion of sites in each bin into account. Most sites have mutation rates on the lower side: bins 1 through 5 contain about 47% of all sites, bins 1 through 19 about 95%, and bins 60 through 97 contain only about 2% of sites. Overall, rates vary 230-fold from lowest to highest. Assuming that the average estimated mutrate of 0.083 corresponds to the genome average mutation rate per site, for which the usual estimate of θ from pairwise differences is about $1/1300 \sim 0.00077$, 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.

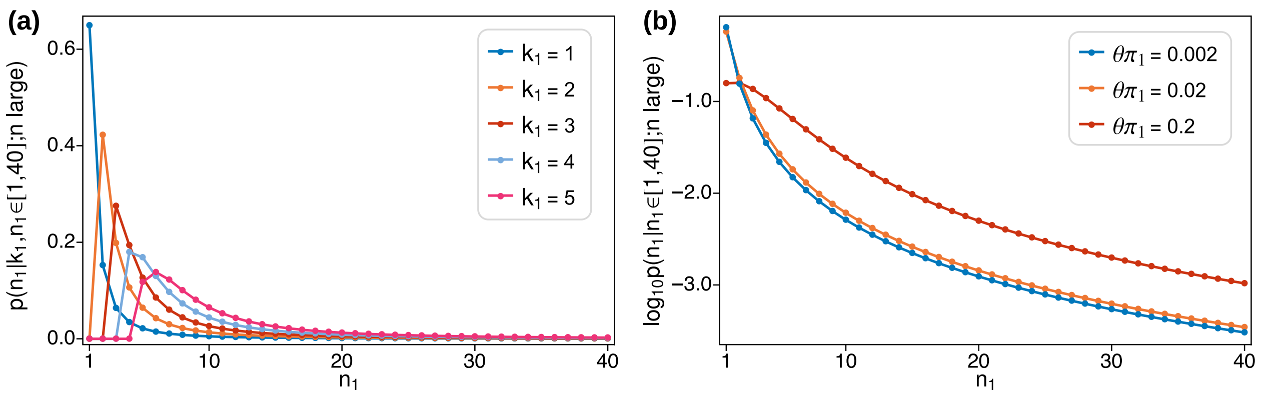

Fig. 3. Plots of the same quantities shown in Fig. 2 but for a sample of size $n = 10^5$ under pure exponential growth with $\beta/n = 3$. a) Probability of observing n_1 copies of allele 1 in the sample given $k_1 = 1, 2, 3, 4, 5$ latent mutations. b) $\log_{10}$-probability of observing n_1 copies 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 Fig. 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 $n_1 \in \{1, 2, \ldots, 40\}$.

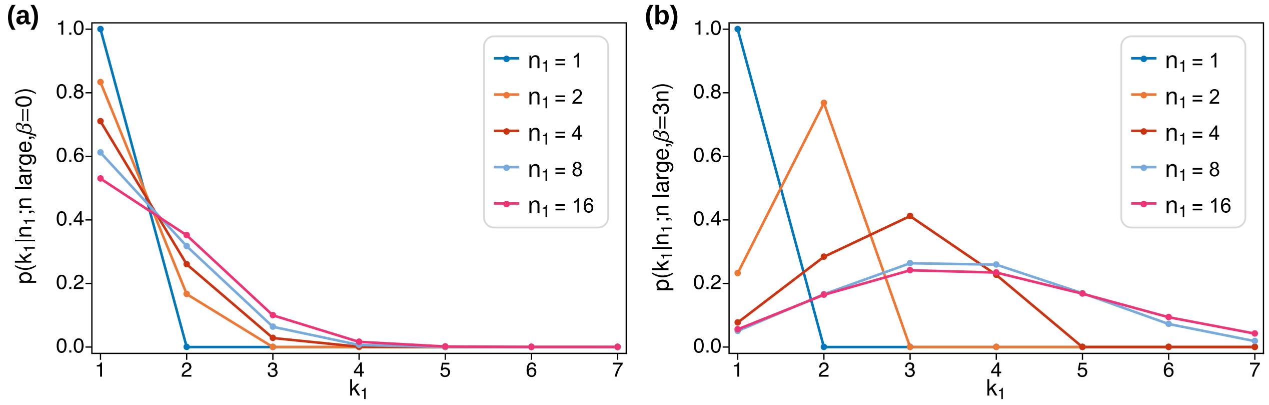

Fig. 4. Probabilities of $k_1 = 1, 2, 3, 4, 5, 6, 7$ latent mutations for increasing values of n_1—1, 2, 4, 8, and 16—when 2.4 mutations are expected on the gene genealogy of a sample of size $n = 10^5$ (or equivalently $\theta\pi_1 = 0.2$ in the constant-size case). Panel A plots (21) with $\theta\pi_1 = 0.2$. Panel B shows the same probability computed using (30) and (31) under exponential growth with $\beta/n = 3$.

We compared observed and expected site-frequency counts for each bin based on an empirical fit of our model. First, we used (30) with the estimated mutrate $(\theta\pi_1/2)\sum_i \bar{\tau}_i$ for each bin to compute probabilities of $k \in \{0, 1, \ldots, 7\}$ latent mutations. Then from (34) and the fact the polymorphisms at sites with very low mutation rates likely have just one latent mutation, we used the combined data for bins 1 through 5 to estimate $\bar{\tau}_i/\sum_i \bar{\tau}_i$ directly as $S_i/(L - S_0)$ for $i \in \{1, \ldots, 40\}$. Our estimates of the mutrate for bins 1-5 range from 0.0097 to 0.037 with an average of 0.021, which we note is somewhat less than the smallest mutation rate in Figures 2 and 3. We assumed that this $\bar{\tau}_i/\sum_i \bar{\tau}_i$ estimated from bins 1–5 holds for all bins. Finally, we computed the expectations $\mathbb{E}[S_i]$, for $i \in \{0, \ldots, 40\}$ in each bin, multiplying the probabilities of counts obtained using (30) and (31) by the total number of sites in the bin, cf. (53).

The upper three panels of Fig. 6 show 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 A2 in the Appendix gives the plots for all 97 bins. 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 depicted in Fig. 6B and in the 50th panel of Fig. A2. The mutrate values in these plots are again the estimates of the expected number of mutations per site on the gene genealogy, $(\theta\pi_1/2)\sum_i \bar{\tau}_i$, for each bin.

The broad pattern from these plots is clear. For smaller mutation rates (e.g. Fig. 6a) the site-frequency spectrum is heavily weighted toward the rarest variants. For large mutation rates (e.g. Fig. 6c), that is when multiple latent mutations are likely, the site-frequency spectrum is shifted toward higher counts. Again from Fig. 5a, the data contain fewer sites with intermediate mutation rates. In this case (e.g. Fig. 6b), the site-frequency spectrum does show 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 Fig. 6a and the first 20 or so panels of Fig. A2, 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. Fig. 6c). 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

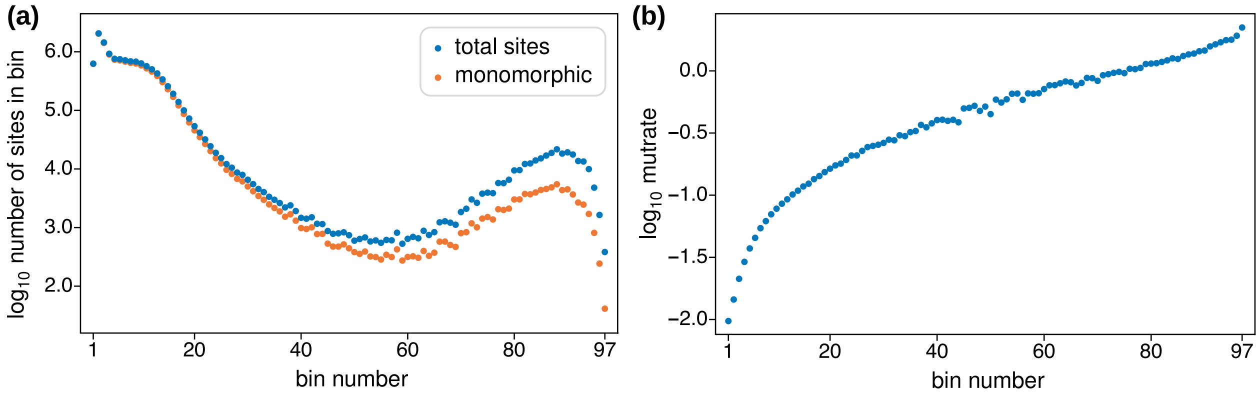

Fig. 5. a) Total numbers of sites and total numbers of monomorphic, or invariant, sites in the gnomAD data for each of the 97 bins. b) Estimated mutation rates—i.e. the “mutrate” or expected number of latent mutations $(\theta\pi_1/2)\sum_i \bar{\tau}_i$ as discussed in the text—on a log scale for bins 1 through 97.

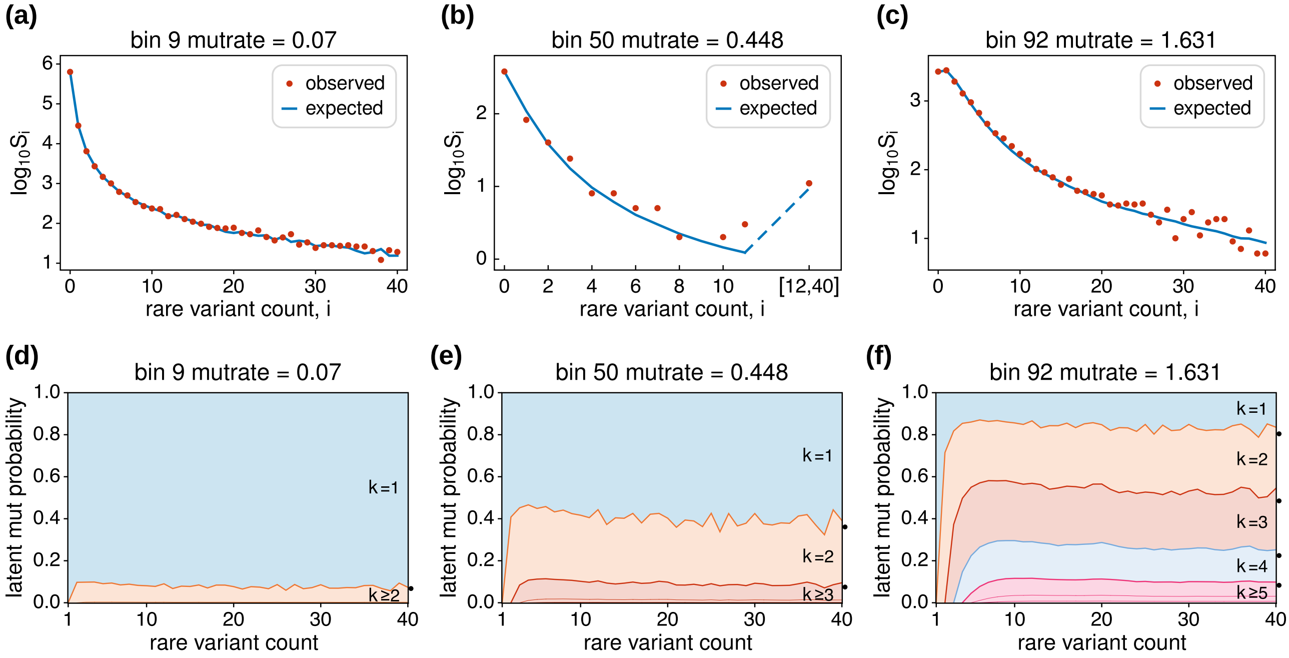

Fig. 6. Upper three panels: Examples of model fit for a) a low-rate bin, b) a middle-rate bin, and c) a high-rate bin. Lower three panels: Stacked probabilities of $k \in \{1, 2, 3, 4, 5, 6, 7\}$ latent mutations for rare variants with counts $i \in \{1, 2, \dots, 20\}$ for the same three bins. As in Fig. 5, “mutrate” indicates an estimate of the expected number of latent mutations per site, $(\theta\pi_1/2)\sum_i \bar{\tau}_i$. Black dots on the right in the lower three panels show the probabilities for the shifted-Poisson result discussed in the text.

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.

The lower three panels of Fig. 6 show estimates of the probability that a variant in count $i \in \{1, \dots, 40\}$ descends from $k \in \{1, \dots, 7\}$ latent mutations, computed using (32). All singletons descend from single mutations. Variants in larger counts can have multiple latent mutations, and the probabilities of these increase very quickly then settle down to stable values. This suggestion of a limiting distribution was also seen for exponential growth in Fig. 4b, only there depicted differently. For very large counts of the variant, the distribution of $k-1$ is well approximated by a Poisson with mean equal to the expected number of mutations per site on the gene genealogy, $(\theta\pi_1/2)\sum_i \bar{\tau}_i$. This shifted-Poisson result is known already for the constant-size case (Arratia *et al.* 2000; Yamato 2017). In “A remark on the total number of mutations for large n_1” in the Appendix we argue that it should hold more generally. The accuracy of this shifted-Poisson result for the gnomAD data and $i = 40$ is shown by the black dots on the right axes of Figs. 6d–f.

For low-rate sites (e.g. Fig. 6d) there is a relatively small chance of multiple latent mutations. But the chance of two or more latent mutations is not negligible, owing to the very large sample size. Note that the mutrate for bin 9 is less than the genome average, which is 0.083 for this sample of $n \sim 114$K. Thus in a very large sample even low-rate sites are affected by recurrent mutation. For the middle-rate sites (e.g. Fig. 6e) in the trough in Fig. 5a the chance of there being only one latent mutation is still considerable. However, for high-rate sites (e.g. Fig. 6f) it can be more likely that there are two or three mutations in the ancestry of a rare variant than the single unique mutation which is typically supposed.

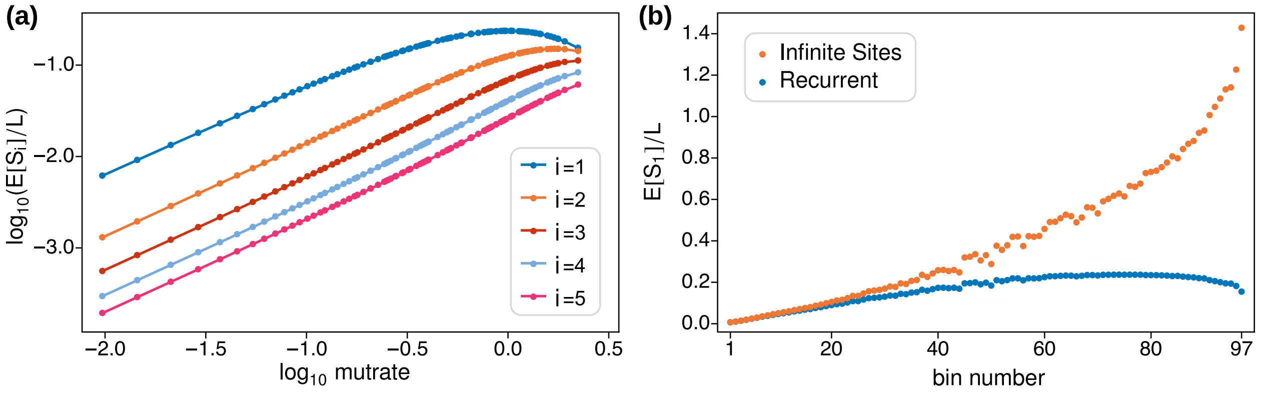

Fig. 7. a) Predicted frequencies of rare variants as a function of mutrate across the 97 bins. $E[S_i]/L$ is the expected proportion of sites at which the rare variant is found in $i \in \{1, 2, \ldots, 5\}$ copies in the sample. The human genome average mutation rate is −1.08 on this scale. b) Predicted frequencies of singletons $E[S_1]/L$ in each bin under the infinite sites mutation model and under the independent-Poissons model of recurrent mutation.

Finally, we explored the extent to which rare variants might be observed less frequently than would be expected if there were no recurrent mutation. Figure 7a shows the expected frequency of singletons, doubletons, etc., up to variants found in five copies in the sample, across the range of mutrates in the binned gnomAD data. The standard infinite-sites prediction is that the frequency will increase linearly with the mutation rate. Figure 7a is largely consistent with this but shows marked deviations when the mutrate becomes too large. The point at which the linear prediction fails depends on the count of the rare variant. Singletons are the first to deviate, which they do as soon as there is an appreciable chance of two or more mutations at a site. For rare variants in five copies, linearity holds even close the upper limit of mutation rates in the human genome.

Figure 7b shows the extent to which the infinite-sites model over-predicts the frequency of singletons across the 97 bins. The infinite-sites prediction for a bin is its mutrate $(\theta\pi_1/2)\sum_i \bar{\tau}_i$ times the proportion of singleton branches $\bar{\tau}_1/\sum_i \bar{\tau}_i = 0.64$ estimated from the first five bins. The corresponding independent-Poissons predictions are the same as those for $i = 1$ in the 97 panels of Fig. A2. The infinite-sites model makes reasonable predictions for the twenty lowest-rate bins, which contain 96% of all sites and have mutation rates less than twice the genome average. But it predicts the impossible for the seven highest-rate bins: more singletons than there are sites to mutate. For bins 21 through 97, which contain 4% of all sites, the infinite-sites model predicts a total of 269,222 singletons compared to the 83,002 which are actually observed.

We emphasize that the results in Fig. 7 depend on the sample size. The expected number of mutations at a single site, $(\theta\pi_1/2)\sum_i \bar{\tau}_i$, is proportional to the total length of the gene genealogy, which is an increasing function of the sample size. Already for the sample size $n \sim$ 114K considered here, singletons start to be affected by recurrent mutation at around the genome average mutation rate (Figs. 7 and 6d). For variants in any fixed count i there will be a sample size above which the infinite-sites, linear prediction starts to fail.

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. In particular, the number of latent mutations is distributed like the number of alleles in the Ewens sampling formula. We obtained more general results, for changing population size, by modeling latent mutations as independent Poisson random variates.

Our aim was to describe how the site-frequency spectra of rare variants in large samples are affected by recurrent mutation. The key parameters for a variant in count i are its expected total rate of mutation on the gene genealogy of the sample (here denoted $(\theta\pi_1/2)\sum_i \bar{\tau}_i$ and called “mutrate” in the previous section) and the expected relative lengths of branches in the gene genealogy which have i descendants in the sample $(\bar{\tau}_i/\sum_i \bar{\tau}_i)$. 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-Finnish 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 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 myriad 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 common allele will be the ancestral source of mutations 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 described our general results as being for populations which may have changed in size. This is appropriate for the general coalescent model (Griffiths and Tavaré 1998) which we assumed for some proofs in the Appendix. Strictly speaking, the general coalescent does not require a generative model for the times between coalescent events. Thus our results can be applied more broadly. The case of a fixed tree with arbitrary τ_i considered in the Appendix is one example. The independent-Poissons model, with results (27) to (35), does not even require interpretation in terms of coalescence times. These results hold 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. Rates of production of mutants have been obtained for under a range of demographies and some types of selection (Lange and Fan 1997; Dorman *et al.* 2004; Lambert 2011; Kaj and Mugal 2016; Torres *et al.* 2020; Müller *et al.* 2022). Applications to selection will likely require free recombination between sites. Desai and Plotkin (2008) applied the independent-Poissons model (for all variant counts in the sample) for example under a version of the Poisson Random Field model (Sawyer and Hartl 1992).

Data availability

The data application to low-frequency synonymous polymorphisms used allele frequencies from exome sequencing data compiled in gnomAD v2.1.1, available here: https://gnomad.broadinstitute.org/downloads and basepair-resolution mutation rates (Seplyarskiy *et al.* 2022), available here: http://genetics.bwh.harvard.edu/downloads/Vova/Roulette/. The mutation rate model specifies the rate for all three possible alternative nucleotides, and different nucleotide mutations were counted separately when generating the site-frequency spectra. The pipeline used to compile and annotate all potential synonymous mutations in the human genome is available at: https://github.com/vseplyarskiy/Roulette. The site-frequency spectra in different mutation rate bins is available at: https://doi.org/10.6084/m9.figshare.3426251.v1.

Acknowledgements

We thank Vladimir Seplyarskiy for helpful discussions and assistance with the mutation-rate binned SFS data from gnomAD, and three anonymous reviewers for helpful comments.

Funding

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 W.-T.(L.)F.; and by National Institutes of Health grants R35-GM127131, R01-MH101244, U01-HG012009 and R01-HG010372 to S.S.

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

Abramowitz M, Stegun IA. Handbook of Mathematical Functions. New York: Dover; 1964.

Achaz G. Frequency spectrum neutrality tests: one for all and all for one. Genetics. 2009;183:249–258. doi:10.1534/genetics.109.104042

Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet. 2016;48:349–355. doi:10.1038/ng.3511

Arratia R, Barbour AD, Tavaré S. Poisson process approximations for the Ewens sampling formula. Ann Appl Probab. 1992;2:519–535. doi:10.1214/aoap/1177005647

Arratia R, Barbour AD, Tavaré S. The number of components in a logarithmic combinatorial structure. Ann Appl Probab. 2000;10: 331–361. doi:10.1214/aoap/1019487347

Arratia R, Barbour AD, Tavaré S. Logarithmic Combinatorial Structures: A Probabilistic Approach. Zürich: European Mathematical Society; 2003 (EMS monographs in mathematics).

Arratia R, Barbour AD, Tavaré S. Exploiting the Feller coupling for the Ewens sampling formula. Stat Sci. 2016;31:27–29. doi:10.1214/15-STS537

Arratia R, Tavaré S. Limit theorems for combinatorial structures via discrete process approximations. Random Struct Algorithms. 1992;3:321–345. doi:10.1002/rsa.3240030310

Baake E, Bialowons R. Ancestral processes with selection: branching and Moran models. Banach Cent Publ. 2008;80:33–52.

Bertoin J. Random Fragmentation and Coagulation Processes. Cambridge: Cambridge University Press; 2006 (Cambridge Studies in Advanced Mathematics.

Bhaskar A, Kamm JA, Song YS. Approximate sampling formulae for general finite-alleles models of mutation. Adv Appl Probab. 2012;44:408–428. doi:10.1239/aap/1339878718

Bhaskar A, Song YS. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. Ann Stat. 2014;42:2469–2493. doi:10.1214/14-AOS1264

Billingsley P. Probability and Measure. New York: John Wiley & Sons; 2008.

Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 1980;8:1499–1504. doi:10.1093/nar/8.7.1499

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 1995;140:783–796. doi:10.1093/genetics/140.2.783

Burden CJ, Griffiths RC. The stationary distribution of a sample from the Wright-Fisher diffusion model with general small mutation rates. J Math Biol. 2019;78:1211–1224. doi:10.1007/s00285-018-1306-y

Burden CJ, Tang Y. An approximate stationary solution for multi-allele neutral diffusion with low mutation rates. Theor Popul Biol. 2016;112:22–32. doi:10.1016/j.tpb.2016.07.005

Burden CJ, Tang Y. Rate matrix estimation from site frequency data. Theor Popul Biol. 2017;113:23–33. doi:10.1016/j.tpb.2016.10.001

Bustamante CD, Wakeley J, Sawyer S, Hartl DL. Directional selection and the site-frequency spectrum. Genetics. 2001;159:1779–1788. doi:10.1093/genetics/159.4.1779

Champagnat N, Lambert A. Splitting trees with neutral Poissonian mutations I: small families. Stoch Process Their Appl. 2012;122: 1003–1033. doi:10.1016/j.spa.2011.11.002

Champagnat N, Lambert A. Splitting trees with neutral Poissonian mutations II: largest and oldest families. Stoch Process Their Appl. 2013;123:1368–1414. doi:10.1016/j.spa.2012.11.013

Cheek D, Antal T. Mutation frequencies in a birth-death branching process. Ann Appl Probab. 2018;28:3922–3947. doi:10.1214/18-AAP1413

Chen H, Chen K. Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. Genetics. 2013;194:721–736. doi:10.1534/genetics.113.151522

Crespo FF, Posada D, Wiuf C. Coalescent models derived from birth-death processes. Theor Popul Biol. 2021;142:1–11. doi:10.1016/j.tpb.2021.09.003

Desai MM, Plotkin JB. The polymorphism frequency spectrum of finitely many sites under selection. Genetics. 2008;180: 2175–2191. doi:10.1534/genetics.108.087361

Donnelly P. Dual processes in population genetics. In: Tautu P, editor. Stochastic Spatial Processes. Berlin: Springer Berlin Heidelberg; 1986. p. 94–105.

Donnelly P, Tavaré S. The population genealogy of the infinitely-many neutral alleles model. J Math Biol. 1987;25:381–391. doi:10.1007/BF00277163

Dorman KS, Sinsheimer JS, Lange K. In the garden of branching processes. SIAM Rev. 2004;46:202–229. doi:10.1137/S0036144502417843

Durrett R. Population genetics of neutral mutations in exponentially growing cancer cell populations. Ann Appl Probab. 2013;23: 230–250. doi:10.1214/11-AAP824

Durrett R. Branching Process Models of Cancer. Cham: Springer; 2015 (Mathematical Biosciences Institute Lecture Series, 1.1).

Eldon B, Birkner M, Blath J, Freund F. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? Genetics. 2015;199:841–856. doi:10.1534/genetics.114.173807

Ewens WJ. The sampling theory of selectively neutral alleles. Theor Popul Biol. 1972;3:87–112. doi:10.1016/0040-5809(72)90035-4

Ewens WJ. A note on the sampling theory for infinite alleles and infinite sites models. Theor Popul Biol. 1974;6:143–148. doi:10.1016/0040-5809(74)90020-3

Ewens WJ. Mathematical Population Genetics. Berlin: Springer-Verlag; 1979.

Ewens WJ. Mathematical Population Genetics, Volume I: Theoretical Foundations. Berlin: Springer-Verlag; 2004.

Fearnhead P. Perfect simulation from population genetic models with selection. Theor Popul Biol. 2001;59:263–279. doi:10.1006/tpbi.2001.1514

Fearnhead P. The common ancestor at a nonneutral locus. J Appl Probab. 2002;39:38–54. doi:10.1017/S0021900200021495

Ferretti L, Ledda A, Wiehe T, Achaz G, Ramos-Onsins SE. Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. Genetics. 2017;207:229–240. doi:10.1534/genetics.116.188763

Fisher RA. The possible modification of the response of the wild type to recurrent mutations. Am Nat. 1928;62:115–126. doi:10.1086/280193

Fisher RA. The distribution of gene ratios for rare mutations. Proc R Soc Edinb. 1930a;50:205–220.

Fisher RA. The Genetical Theory of Natural Selection. Oxford: Clarendon; 1930b.

Fisher RA. A theoretical distribution for the apparent abundance of different species. J Anim Ecol. 1943;12:54–57.

Fu Y. Statistical properties of segregating sites. Theor Popul Biol. 1995;48:172–197. doi:10.1006/tpbi.1995.1025

Gao F, Keinan A. Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. Genetics. 2016;202:235–245. doi:10.1534/genetics.115.180570

Gazave E, Ma L, Chang D, Coventry A, Gao F, Muzny D, Boerwinkle E, Gibbs R, Sing CF, Clark AG, *et al.* Neutral genomic regions refine models of recent rapid human population growth. Proc Natl Acad Sci USA. 2014;111:757–762. doi:10.1073/pnas.1310398110

Goldman N. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. Nucleic Acids Res. 1993;21:2487–2491. doi:10.1093/nar/21.10.2487

Griffiths RC. Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theor Popul Biol. 1980;17:37–50. doi:10.1016/0040-5809(80)90013-1

Griffiths RC, Tavaré S. Ancestral inference in population genetics. Stat Sci. 1994a;9:307–319. doi:10.1214/ss/1177010378

Griffiths RC, Tavaré S. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B: Biol Sci. 1994b;344: 403–410. doi:10.1098/rstb.1994.0079

Griffiths RC, Tavaré S. The age of a mutation in a general coalescent tree. Commun Stat Stoch Models. 1998;14:273–295. doi:10.1080/15326349808807471

Gunnarsson EB, Leder K, Foo J. Exact site frequency spectra of neutrally evolving tumors: a transition between power laws reveals a signature of cell viability. Theor Popul Biol. 2021;142:67–90. doi:10.1016/j.tpb.2021.09.004

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5: 1–11. doi:10.1371/journal.pgen.1000695

Haldane JBS. The part played by recurrent mutation in evolution. Am Nat. 1933;67:5–19. doi:10.1086/280465

Harpak A, Bhaskar A, Pritchard JK. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. PLoS Genet. 2016;12:e1006489. doi:10.1371/journal.pgen.1006489

Hobolth A, Wiuf C. The genealogy, site frequency spectrum and ages of two nested mutant alleles. Theor Popul Biol. 2009;75:260–265. Sam Karlin: Special Issue. doi:10.1016/j.tpb.2009.02.001

Hudson RR. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 1983;37:203–217. doi:10.2307/2408186

Jenkins PA, Mueller JW, Song YS. General triallelic frequency spectrum under demographic models with variable population size. Genetics. 2014;196:295–311. doi:10.1534/genetics.113.158584

Jenkins PA, Song YS. The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. Theor Popul Biol. 2011;80:158–173. doi:10.1016/j.tpb.2011.04.001

Johnson KE, Adams CJ, Voight BF. Identifying rare variants inconsistent with identity-by-descent in population-scale whole-genome sequencing data. Methods Ecol Evol. 2022;13:2429–2442. doi:10.1111/2041-210X.13991

Kaj I, Mugal CF. The non-equilibrium allele frequency spectrum in a poisson random field framework. Theor Popul Biol. 2016;111: 51–64. doi:10.1016/j.tpb.2016.06.003

Kaplan N, Hudson RR. The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. Theor Popul Biol. 1985;28:382–396. doi:10.1016/0040-5809(85)90036-X

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–443. doi:10.1038/s41586-020-2308-7

Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012;336: 740–743. doi:10.1126/science.1217283

Kern AD, Hey J. Exact calculation of the joint allele frequency spectrum for isolation with migration models. Genetics. 2017;207: 241–253. doi:10.1534/genetics.116.194019

Kessler DA, Levine H. Large population solution of the stochastic Luria & Delbrück evolution model. Proc Natl Acad Sci USA. 2013;110:11682–11687. doi:10.1073/pnas.1309667110

Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. Genetics. 1969;61:893–903. doi:10.1093/genetics/61.4.893

Kimura M. Theoretical foundation of population genetics at the molecular level. Theor Popul Biol. 1971;2:174–208. doi:10.1016/0040-5809(71)90014-1

Kingman JFC. On the genealogy of large populations. J Appl Probab. 1982;19:27–43. doi:10.1017/S0021900200034446

Lambert A. Species abundance distributions in neutral models with immigration or mutation and general lifetimes. J Math Biol. 2011; 63:57–72. doi:10.1007/s00285-010-0361-9

Lange K, Fan Rz. Branching process models for mutant genes in nonstationary populations. Theor Popul Biol. 1997;51:118–133. doi:10.1006/tpbi.1997.1297

Lapierre M, Lambert A, Achaz G. Accuracy of demographic inferences from the site frequency spectrum: the case of the Yoruba population. Genetics. 2017;206:439–449. doi:10.1534/genetics.116.192708

Lea DE, Coulson A. The distribution of the numbers of mutants in bacterial populations. J Genet. 1949;49:264–285. doi:10.1007/BF02986080

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Séurel L, Venkat A, Andolfatto P, Przeworski M. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 2012;10: 1–9. doi:10.1371/journal.pbio.1001388

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, *et al.* Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–291. doi:10.1038/nature19057

Liu X, Fu YX. Exploring population size changes using snp frequency spectra. Nat Genet. 2015;47:555–559. doi:10.1038/ng.3254

Luria SE, Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. Genetics. 1943;28:491–511. doi:10.1093/genetics/28.6.491

Moran PAP. Random processes in genetics. Proc Camb Phil Soc. 1958; 54:60–71. doi:10.1017/S0305004100033193

Moran PAP. Statistical Processes of Evolutionary Theory. Oxford: Clarendon Press; 1962.

Müller R, Kaj I, Mugal CF. A nearly neutral model of molecular signatures of natural selection after change in population size. Genome Biol Evol. 2022;14:evac058.

Myers S, Fefferman C, Patterson N. Can one learn history from the allelic spectrum? Theor Popul Biol. 2008;73:342–348. doi:10.1016/j.tpb.2008.01.001

Nielsen R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics. 2000; 154:931–942. doi:10.1093/genetics/154.2.931

Ohtsuki H, Innan H. Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. Theor Popul Biol. 2017;117:43–50. doi:10.1016/j.tpb.2017.08.006

Polanski A, Kimmel M. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics. 2003;165: 427–436. doi:10.1093/genetics/165.1.427

Polanski A, Szczesna A, Garbulowski M, Kimmel M. Coalescence computations for large samples drawn from populations of time-varying sizes. PLoS ONE. 2017;12:1–22. doi:10.1371/journal.pone.0170701

Poon GYP, Watson CJ, Fisher DS, Blundell JR. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. Nat Genet. 2021;53:1597–1605. doi:10.1038/s41588-021-00957-1

Rannala B, Slatkin M. Estimating the age of alleles by use of intraallelic variability. Am J Human Genet. 1997;60:447–458.

Rosen Z, Bhaskar A, Roch S, Song YS. Geometry of the sample frequency spectrum and the perils of demographic inference. Genetics. 2018;210:665–682. doi:10.1534/genetics.118.300733

Sargsyan O. 2006. Analytical and simulation results for the general coalescent [PhD thesis]. Los Angeles: University of Southern California.

Sargsyan O. An analytical framework in the general coalescent tree setting for analyzing polymorphisms created by two mutations. J Math Biol. 2015;70:913–956. doi:10.1007/s00285-014-0785-8

Saunders IW, Tavaré S, Watterson GA. On the genealogy of nested subsamples from a haploid population. Adv Appl Probab. 1984; 16:471–491. doi:10.2307/1427285

Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. Genetics. 1992;132:1161–1176. doi:10.1093/genetics/132.4.1161

Schrempf D, Hobolth A. 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. Theor Popul Biol. 2017;114:88–94. doi:10.1016/j.tpb.2016.12.001

Seplyarskiy V, Lee DJ, Koch EM, Lichtman JS, Luan HH, Sunyaev SR. A mutation rate model at the basepair resolution identifies the mutagenic effect of Polymerase III transcription. bioRxiv 504670. https://doi.org/10.1101/2022.08.20.504670, 2022, preprint: not peer reviewed.

Seplyarskiy VB, Soldatov RA, Koch E, McGinty RJ, Goldmann JM, Hernandez RD, Barnes K, Correa A, Burchard EG, Ellinor PT, *et al.* Population sequencing data reveal a compendium of mutational processes in the human germ line. Science. 2021;373: 1030–1035. doi:10.1126/science.aba7408

Slade PF. Most recent common ancestor probability distributions in gene genealogies under selection. Theor Popul Biol. 2000a;58: 291–305. doi:10.1006/tpbi.2000.1488

Slade PF. Simulation of selected genealogies. Theor Popul Biol. 2000b; 57:35–49. doi:10.1006/tpbi.1999.1438

Slatkin M. Allele age and a test for selection on rare alleles. Philos Trans R Soc Lond B: Biol Sci. 2000;355:1663–1668. doi:10.1098/rstb.2000.0729

Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 1991;129:555–562. doi:10.1093/genetics/129.2.555

Slatkin M, Rannala B. Estimating allele age. Annu Rev Genomics Hum Genet. 2000;1:225–249. doi:10.1146/annurev.genom.1.1.225

Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics. 2009;182:205–216.

Stephens M, Donnelly P. Ancestral inference in population genetics models with selection (with discussion). Aust N Z J Stat. 2003; 45:395–430. doi:10.1111/1467-842X.00295

Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105:437–460. doi:10.1093/genetics/105.2.437

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

Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590:290–299. doi:10.1038/s41586-021-03205-y

Terhorst J, Song YS. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. Proc Natl Acad Sci USA. 2015;112:7677–7682. doi:10.1073/pnas.1503717112

The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.

Torres R, Stetter MG, Hernandez RD, Ross-Ibarra J. The temporal dynamics of background selection in nonequilibrium populations. Genetics. 2020;214:1019–1030. doi:10.1534/genetics.119.302892

Tricomi F, Erdélyi A. The asymptotic expansion of a ratio of gamma functions. Pac J Appl Math. 1951;1:133–142. doi:10.2140/pjm.1951.1.133

Vogl C, Clemente F. The allele-frequency spectrum in a decoupled Moran model with mutation, drift, and directional selection, assuming small mutation rates. Theor Popul Biol. 2012;81: 197–209. doi:10.1016/j.tpb.2012.01.001

Vogl C, Mikula LC, Burden CJ. Maximum likelihood estimators for scaled mutation rates in an equilibrium mutation-drift model. Theor Popul Biol. 2020;134:106–118. doi:10.1016/j.tpb.2020.06.001

Watterson GA. Models for the logarithmic species abundance distributions. Theor Popul Biol. 1974a;6:217–250. doi:10.1016/0040-5809(74)90025-2

Watterson GA. The sampling theory of selectively neutral alleles. Adv Appl Probab. 1974b;6:463–488. doi:10.2307/1426228

Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7:256–276. doi:10.1016/0040-5809(75)90020-9

Watterson GA. Lines of descent and the coalescent. Theor Popul Biol. 1984;26:77–92. doi:10.1016/0040-5809(84)90025-X

Wiuf C. On the genealogy of a sample of neutral rare alleles. Theor Popul Biol. 2000;58:61–75. doi:10.1006/tpbi.2000.1469

Wiuf C, Donnelly P. Conditional genealogies and the age of a neutral mutant. Theor Popul Biol. 1999;56:183–201. doi:10.1006/tpbi.1998.1411

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

Wright S. Evolution in Mendelian populations. Genetics. 1931;16: 97–159. doi:10.1093/genetics/16.2.97

Wright S. The distribution of gene frequencies under irreversible mutation. Proc Natl Acad Sci USA. 1938;24:253–259. doi:10.1073/pnas.24.7.253

Wright S. Adaptation and selection. In: Jepson GL, Simpson GG, Mayr E, editors. Genetics, Paleontology and Evolution. Princeton (RI): Princeton University Press; 1949.

Yamato H. Poisson approximations for sum of Bernoulli random variables and its application to Ewens sampling formula. J Jpn Stat Soc. 2017;47:187–195. doi:10.14490/jjss.47.187

Appendix

Time-dependent conditional ancestral process

Here we study the conditional ancestral process in detail and provide the justification for (18) and (19).

Let $\mathcal{N}_1(t)$ and $\mathcal{N}_2(t)$ be the numbers of rare alleles and common alleles respectively at time t. From (17a), (17b) and (17c), 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

$$(n_1, n_2) \to \begin{cases} (n_1-1, n_2+1) & \text{w/prob. } \frac{\theta\pi_1}{\theta\pi_1+n_1-1}\frac{n_1}{n_2} \\ (n_1-1, n_2) & \text{w/prob. } \frac{n_1-1}{\theta\pi_1+n_1-1}\frac{n_1}{n_2} \\ (n_1, n_2-1) & \text{w/prob. } 1-\frac{n_1}{n_2} \end{cases} \tag{A1}$$

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

$$\mathcal{T}_i := \inf\{t \ge 0 : \mathcal{N}_1(t) = n_1 - i\} \tag{A2}$$

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 $(n_1, n_2) \to (n_1 - 1, \mathcal{N}_2(\mathcal{T}_1)) \to \cdots \to (0, \mathcal{N}_2(\mathcal{T}_{n_1}))$.

In Theorem A1 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 A1 As $n_2 \to \infty$, the random vector

$$\left(n_2(\mathcal{T}_i - \mathcal{T}_{i-1}),\ \frac{\mathcal{N}_2(\mathcal{T}_i)}{n_2}\right)_{i=1}^{n_1} \tag{A3}$$

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

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

Remark A1 (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 A1 implies that $\mathbb{E}_{\mathbf{n}}[\mathcal{T}_1]$ is of order $1/n_2$ and gives the second part of (18) 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 $k \in \{0, 1, \ldots n_2 - 1\}$ by (A6). Hence by (A4) and Fubini'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 (18) in the main text.

Remark A2 (Mean of $\mathcal{N}_2(\mathcal{T}_1)$). By (A5) and Theorem A1,

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

for $1 \le i \le n_1$. This gives (19) in the main text.

Proof of Theorem A1

To explain the key idea we first establish weak convergence of $(n_2\mathcal{T}_1, \frac{\mathcal{N}_2(\mathcal{T}_1)}{n_2})$, i.e. of the marginal distribution for $i = 1$ in (A3). By definition, $\mathcal{T}_1$ is given by

$$\mathcal{T}_1 = \sum_{i=0}^{\sharp}\xi_i, \tag{A4}$$

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

$$\begin{cases} n_2-\sharp+1, & \text{w/conditional prob. } \frac{\theta\pi_1}{\theta\pi_1+n_1-1}\frac{n_1}{n_2-\sharp} \\ n_2-\sharp, & \text{w/conditional prob. } \frac{n_1-1}{\theta\pi_1+n_1-1}\frac{n_1}{n_2-\sharp} \end{cases} \tag{A5}$$

which correspond to a non-empty mutation event and a coalescent event of type 1 respectively. These follow from (A1).

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

$$\begin{aligned}
&\mathbb{P}_{\mathbf{n}}(\sharp=k)\\
&\quad=\left(1-\frac{n_1}{n_2}\right)\left(1-\frac{n_1}{n_2-1}\right)\cdots\left(1-\frac{n_1}{n_2-k+1}\right)\frac{n_1}{n_2-k}\\
&\quad=\frac{n_1}{n_2}\prod_{j=1}^{k}\frac{n_2-n_1-j+1}{n_2-j} &&\text{(A6)}\\
&\quad=\frac{n_1}{n_2}\prod_{j=1}^{n_1-1}\frac{n_2-k-j}{n_2-j}\\
&\quad\approx\frac{n_1}{n_2}(1-x)^{n_1-1} &&\text{(A7)}
\end{aligned}$$

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 $k\in\{0, 1, 2, \ldots, n_2-n_1-1\}$,

$$\begin{aligned}
\mathbb{P}_{\mathbf{n}}(\sharp>k)&=\left(1-\frac{n_1}{n_2}\right)\left(1-\frac{n_1}{n_2-1}\right)\cdots\left(1-\frac{n_1}{n_2-k}\right)\\
&\approx\prod_{j=1}^{n_1}\left(1-\frac{k+j}{n_2}\right) &&\text{(A8)}\\
&\approx(1-x)^{n_1}
\end{aligned}$$

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

Lemma A1. As $n_2\to\infty$, we have convergence in distribution

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

with Z_1 and Y_1 as defined in Theorem A1.

Proof of Lemma A1. It suffices to show that the moment generating function of the $\mathbb{R}^2$-valued random variable on the left converges pointwise to that on the right; that is, to show that

$$\lim_{n_2\to\infty}\mathbb{E}_{\mathbf{n}}\left[e^{\eta\frac{\sharp}{n_2}+\zeta n_2\mathcal{T}_1}\right]=n_1\int_0^1(1-x)^{n_1-1}e^{\{\eta x+\frac{2\zeta x}{1-x}\}}\,dx \tag{A9}$$

for $\eta\in\mathbb{R}$ and $\zeta\in(-\infty,0]$. See, for instance, Section 30 of Billingsley (2008). Since $\xi_i\sim\text{Exp}(\lambda(n_1, n_2-i))$,

$$\mathbb{E}_{\mathbf{n}}[e^{\zeta\xi_i}]=\frac{\lambda(n_1,n_2-i)}{\lambda(n_1,n_2-i)-\zeta}=\frac{(n_2-i)^2}{(n_2-i)^2-2\zeta}. \tag{A10}$$

By (A4), (A6) and (A10),

$$\begin{aligned}
\mathbb{E}_{\mathbf{n}}\left[e^{\eta\frac{\sharp}{n_2}+\zeta n_2\mathcal{T}_1}\right]&=\sum_{k=0}^{n_2-n_1}\mathbb{P}_{\mathbf{n}}(\sharp=k)\,e^{\eta\frac{k}{n_2}}\,\mathbb{E}_{\mathbf{n}}\left[e^{\zeta n_2\sum_{i=0}^{k}\xi_i}\right]\\
&=\sum_{k=0}^{n_2-n_1}\mathbb{P}_{\mathbf{n}}(\sharp=k)\,e^{\eta\frac{k}{n_2}}\prod_{i=0}^{k}\mathbb{E}_{\mathbf{n}}[e^{\zeta n_2\xi_i}] &&\text{(A11)}\\
&=\frac{n_1}{n_2}\sum_{k=0}^{n_2-n_1}e^{\eta\frac{k}{n_2}}\prod_{j=1}^{n_1-1}\frac{n_2-k-j}{n_2-j}\,p_{n_2}(\zeta),
\end{aligned}$$

where

$$\begin{aligned}
p_{n_2}(\zeta):&=\prod_{i=0}^{k}\frac{\lambda(n_1,n_2-i)}{\lambda(n_1,n_2-i)-\zeta n_2}\\
&=\exp\left\{-\sum_{i=0}^{k}\log\left(1-\frac{2\zeta n_2}{(n_2-i)^2}\right)\right\}\\
&\approx\exp\left\{2\zeta n_2\sum_{i=0}^{k}\frac{1}{(n_2-i)^2}\right\}\qquad\text{if }\frac{2\zeta n_2}{(n_2-i)^2}\approx 0 &&\text{(A12)}\\
&\approx\exp\left\{2\zeta\int_0^x\frac{1}{(1-y)^2}\,dy\right\}=\exp\left\{\frac{2\zeta x}{1-x}\right\}
\end{aligned}$$

if $\frac{k}{n_2}\to x\in(0,1)$ and $n_2\to\infty$. Putting (A12) and (A7) into (A11), we obtain the desired (A9) and thus Lemma A1. □

We now return to the proof of Theorem A1. Lemma A1 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

$$\lim_{n_2\to\infty}\mathbb{P}_{\mathbf{n}}(\mathcal{N}_2(\mathcal{T}_1)>M)=1\qquad\text{for all }M\in(0,\infty). \tag{A13}$$

As in (A4), 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 (A5), $\mathcal{N}_2(\mathcal{T}_2)$ 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 (A13). 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

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

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 A1. This implies the convergence in distribution

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

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 A1. □

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 "Theory for nonconstant populations." 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. Our definition of nested differs from some previous ones (Saunders *et al.* 1984; Wiuf and Donnelly 1999; Hobolth and Wiuf 2009); here nested mutations may occur on the same branch of the gene genealogy.

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 $k \in \{2, 3, \ldots, n\}$.
2) τ_j for $j \in \{1, \ldots, n-1\}$, 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 ith 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 $i \in \{1, 2, \ldots, m_b\}$. Fig. A1 illustrates this for $b = 4$.

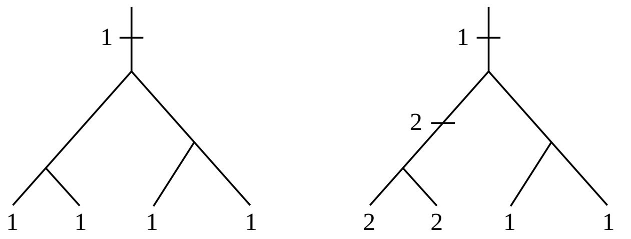

Fig. A1. Two subtrees in $\{\Gamma_i^{(4)}\}$. The subtree on the left has one mutation which is labeled 1 and has count four. The subtree on the right has nested mutations, with the mutation labeled 1 in count two and another labeled 2 also in count two.

We assume that mutations arise as a Poisson point process on the tree with constant rate $\theta/2$ per unit length. Theorem A2 below holds for any fixed ultrametric tree (it can be binary or have multiple mergers, or even be a star tree).

Theorem A2 (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

$$\min\left\{\frac{\theta^2}{8} b\, T_{\text{total}}\, H_n,\ \frac{\theta^2}{8} b^3 \sum_{j=1}^{b} \sum_{k=1}^{m_j} \tau_{j,k}^2\right\}. \tag{A14}$$

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

Remark A3. There is good evidence that the upper 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, $b\theta H_n/4$, 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 "Application to human SNP data."

Remark A4. The simpler bound $\frac{\theta^2}{8} bT_{\text{total}}H_n$ 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 (A14) 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 (A17). 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

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

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,

$$\begin{aligned}
&\mathbb{P}(\text{there are 2 mutations on } \Gamma_i^{(b)} \text{ for some } i \in \{1, 2, \ldots, m_b\}) \\
&\quad \leq \sum_{k=0}^{\infty} e^{-c_n} \frac{c_n^k}{k!} \frac{k(k-1)}{2} \sum_{i=1}^{m_b} \frac{|\Gamma_i^{(b)}|^2}{T_{\text{total}}^2} \\
&\quad = \frac{c_n^2}{2} \sum_{i=1}^{m_b} \frac{|\Gamma_i^{(b)}|^2}{T_{\text{total}}^2} \\
&\quad = \frac{\theta^2}{8} \sum_{i=1}^{m_b} |\Gamma_i^{(b)}|^2.
\end{aligned} \tag{A15}$$

Note that $|\Gamma_i^{(b)}| \leq bH_n$ for all $1 \leq i \leq m_b$, and that $\sum_{i=1}^{m_b} |\Gamma_i^{(b)}| \leq T_{\text{total}}$ since the subtrees $\{\Gamma_i^{(b)}\}_{i=1}^{m_b}$ are disjoint. Hence

$$\sum_{i=1}^{m_b} |\Gamma_i^{(b)}|^2 \leq bH_n \sum_{i=1}^{m_b} |\Gamma_i^{(b)}| \leq bT_{\text{total}}H_n.$$

Putting this into (A15), we obtained the first bound $\frac{\theta^2}{8} bT_{\text{total}}H_n$ in (A14). To get the second bound in (A14), note that $|\Gamma_i^{(b)}| \le bH_i^{(b)}$ for all $1 \le i \le m_b$, 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 $j = b, b-1, \dots, 2, 1$, and these branches are pairwise disjoint for different i's (for $1 \le i \le m_b$). Hence

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

where we used the general inequality $|\sum_{k=1}^b a_k|^2 \le b\sum_{k=1}^b a_k^2$. The bound in (A14) now follows by putting these into (A15). □

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 Fig. A1.

Theorem A3 (Poisson approximation for counts on a fixed tree). Let $\mathbf{T}_n$ be a fixed coalescent tree with n leaves for $n \ge 2$. 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 $a_j \sim \text{Poisson}(\frac{\theta}{2}\tau_j)$ for $1 \le j \le b$.

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 $1 \le j \le b$. 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 $\theta/2$. □

Nested mutation on random trees

We now suppose the tree $\mathbf{T}_n$ is a *random* binary tree (for $n \ge 2$), in particular the general coalescent tree of Griffiths and Tavaré (1998). For each $n \ge 2$, $\{T_k\}_{k=2}^n$ 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), $\lambda(t) = e^{-\beta t}$ with $\beta > 0$ corresponds to an exponentially growing population.

Theorem A4 (Nested mutation on random trees for fixed θ). Let $b \in \mathbb{N}$. Suppose for $1 \le j \le b$,

$$\lim_{n\to\infty} \mathbf{E}_n\left[\sum_{k=1}^{m_j} \tau_{j,k}^2\right] = 0, \tag{A16}$$

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 $\{C_{b,n}\}_{n\ge2}$ are constants that tend to 0 as $n \to \infty$. Furthermore, (A16) holds for the generalized coalescent trees of Griffiths and Tavaré (1998) when $\sup_{t\ge0}\lambda(t) < \infty$ (which includes any growing population).

Proof. The first statement follows directly from Theorem A2. By the fact $\sum_{k=1}^{m_j} \tau_{j,k}^2 \le (\max_{1\le k\le m_j} \tau_{j,k})\tau_j$ and the Cauchy-Schwarz inequality, we have

$$\mathbf{E}_n\left[\sum_{k=1}^{m_j} \tau_{j,k}^2\right] \le \sqrt{\mathbf{E}_n[\tau_j^2]\,\mathbf{E}_n\left[\left(\max_{1\le k\le m_j} \tau_{j,k}\right)^2\right]}. \tag{A17}$$

Hence assumption (A16) is satisfied if

$$\lim_{n\to\infty} \mathbf{E}_n\left[\left(\max_{1\le k\le m_j} \tau_{j,k}\right)^2\right] = 0 \tag{A18}$$

$$\limsup_{n\to\infty} \mathbf{E}_n\left[\tau_j^2\right] < \infty, \tag{A19}$$

for $1 \le j \le b$. The second statement now follows from Lemmas A2, A3, and Proposition A1 below. □

Lemma A2 concerns assumption (A18). For reference, we note that it is satisfied, and hence (A18) 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 A2 Suppose $\limsup_{n\to\infty} \sum_{k=2}^n T_k$ has finite pth moment, where $p > 0$. Then $\max_{1\le k\le m_j} \tau_{j,k} \to 0$ 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.

We construct the infinite sequence $\{\mathbf{T}_n\}_{n\ge2}$ sequentially in the same probability space, by constructing a coupling of the two independent families $\{T_k\}_{k\ge2}$ and $\{\iota_k\}_{k\ge2}$, where $\iota_n \in \{1, 2, \dots, n\}$ is the index of the lineage that branches into two going from $\mathbf{T}_n$ to $\mathbf{T}_{n+1}$.

Let $A_\ell^{(k,n)}$ be the number of descendants in $\mathbf{T}_n$ of the ℓth line of state k. Note that $A_\ell^{(k,n)} \ge 1$ for $\ell \in \{1, 2, \dots, k\}$, and $\sum_{\ell=1}^k A_\ell^{(k,n)} = n$. By exchangeability—in particular see Bertoin (2006, Proposition 2.8)—the random vector $\frac{1}{n}(A_1^{(k,n)}, \dots, A_k^{(k,n)})$ 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,

$$\lim_{n\to\infty} A_\ell^{(k,n)} = +\infty \quad \text{for all } k \ge 1 \text{ and } \ell = 1, 2, \dots, k. \tag{A20}$$

Since $\sum_{k=2}^\infty T_k$ is finite almost surely, the trees $\{\mathbf{T}_n\}_{n\ge2}$ have uniformly bounded height almost surely. So (A20) implies that with probability one,

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

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

Next consider assumption (A19). For the Kingman coalescent, τ_j is close to its mean $\mathbf{E}_n[\tau_j] = 2/j$ because for n large enough,

$$\text{Var}(\tau_j) = 4\sigma_{jj} \le \frac{4(j+1)\log n}{n}, \tag{A21}$$

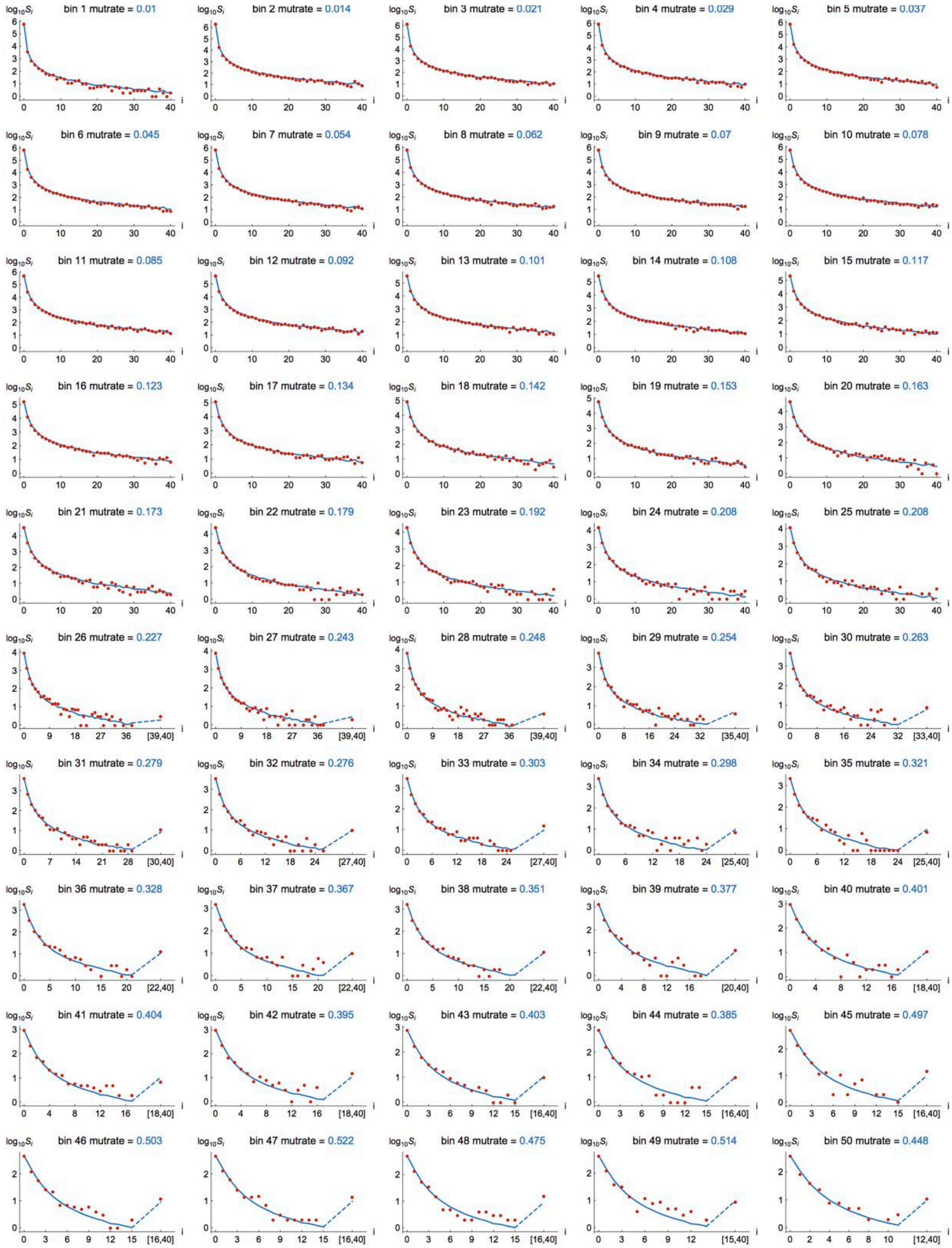

Fig. A2. Plots like those in Fig. 6 for each of the 97 mutation-rate bins. (continues on next page)

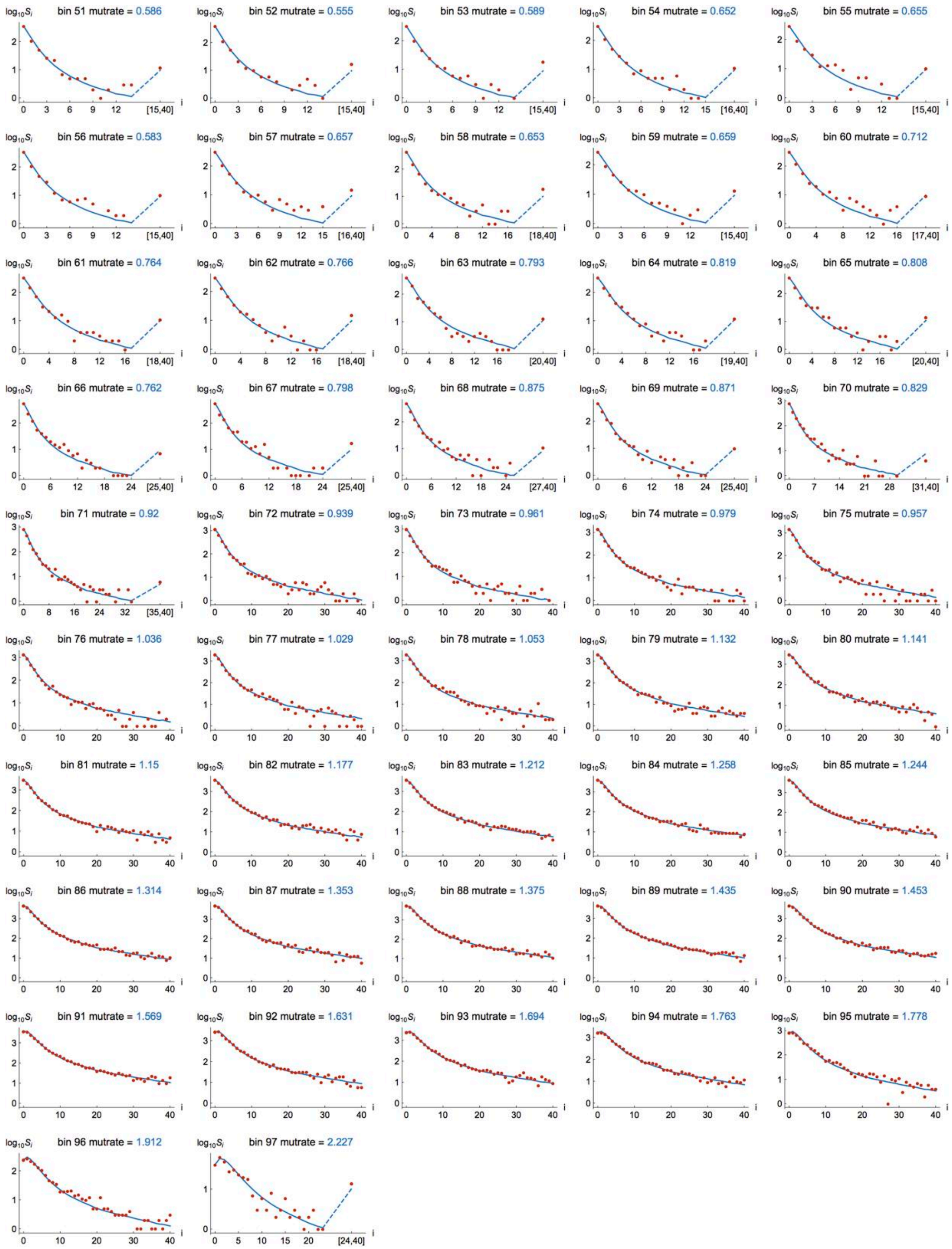

Fig. A2. Continued

where σ_{jj} is defined in Fu (1995, eqns. (1)-(2)). This follows from the fact that Fu's $\beta_n(j) \approx \frac{2\log n}{n}$ 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 A3. Suppose there exists a constant $C_* \in (0, \infty)$ such that

$$\sup_{n\geq 2} \mathbf{E}_n[T_k^2] \leq \frac{C_*}{k^4} \qquad \text{for all } k \geq 2. \tag{A22}$$

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

Proof. For realized values of T_k, the argument in Fu (1995, p. 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 ℓth line of state k defined in the proof of Lemma A2.

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

$$\mathbf{E}_n[\tau_j \mid \{T_k\}_{k=2}^n] = \sum_{k=2}^{n} T_k\, k\, p(k, j) \tag{A23}$$

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

$$\begin{aligned}\mathbf{E}_n[\tau_j^2 \mid \{T_k\}_{k=2}^n] = &\sum_{k=2}^{n} T_k^2 (kp(k, j) + k(k-1)p(k, j; k, j)) \\ &+ 2\sum_{k<k'} T_k T_{k'} kk' p(k, j; k', j),\end{aligned} \tag{A24}$$

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 (A23) and (A24). The bound $\mathbf{E}_n[\tau_j] \leq \frac{\sqrt{C_*}}{j}$ follows from the same calculation in Fu (1995, eqn. (22)). By (A24), 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 (A22), $\limsup_{n\to\infty} \mathbf{E}_n[\tau_j^2] < \infty$ holds also for our random trees. □

Remark A5. As in Theorem A2, we can use an alternate assumption than A16. For any positive integer b, the probability that nested mutation up to count b occurs is bounded above by $\frac{b\theta^2}{8} \mathbf{E}_n[T_{\text{total}} H_n]$ which tends to 0 if $\theta^2 \mathbf{E}_n[T_{\text{total}} H_n] \to 0$. For Kingman coalescent trees, this would require that $\theta \to 0$.

We now check that the assumption (A16) in Theorem A4 holds for the generalized coalescent tree of Griffiths and Tavaré (1998).

Proposition A1. 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 A2 (with $p = 2$) and Lemma A3. In particular, (A16) is satisfied and so the conclusion of Theorem A4 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 $k2$ at state $k \in \{1, 2, \ldots, n\}$, starting at $D_n(0) = n$, and let

$$D_n^{(\lambda)}(t) = D_n\left(\int_0^t \frac{1}{\lambda(u)}\, du\right) \tag{A25}$$

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 (A25), the jump times of the pure death process D_n, denoted by $\tilde{\sigma}_1 < \tilde{\sigma}_2 < \cdots < \tilde{\sigma}_{n-1}$, are given by $\int_0^{\sigma_j} \frac{1}{\lambda} = \tilde{\sigma}_j$ for $1 \leq j \leq n-1$. Hence, with the convention $\sigma_0 = 0$, for $0 \leq j \leq n-2$ we have

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

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

Since $\tilde{\sigma}_{n-k+1} - \tilde{\sigma}_{n-k}$ is equal in distribution to the analog 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 (A18) and (A19) are satisfied. □

Replacing τ_j by its mean

By using the expected coalescence times denoted $\bar{\tau}_i$ in the main text, we implicitly assumed that different sites have different trees and that these are all drawn from the same distribution. Theorem A5 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 A3 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 A4.

Lemma A4. Suppose (A22) holds and that the covariance

$$\mathrm{Cov}(T_k, T_{k'}) \leq \frac{C_n}{k(k-1)k'(k'-1)} \tag{A26}$$

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 (A23) and (A24) with respect to $\mathbf{E}_n$, we obtain the variance

$$\begin{aligned}\mathrm{Var}(\tau_j) &= \mathbf{E}_n[\tau_j^2] - (\mathbf{E}_n[\tau_j])^2 \\ &= 2\sum_{k<k'} kk'(\mathbf{E}_n[T_k T_{k'}]p(k, j; k', j) \\ &\quad - \mathbf{E}_n[T_k]\mathbf{E}_n[T_{k'}]p(k, j)p(k', j))\end{aligned} \tag{A27}$$

up to an $O(\frac{\log n}{n})$ term. This follows from Fu (1995, eqns. (24)-(25)) and assumption (A22) in Lemma A3. This also leads to (A21).

By assumptions (A22) and (A26), the double sum in (A27) is bounded above by

$$C_n \sum_{k<k'} \frac{p(k,j;k',j)}{(k-1)(k'-1)} + C_* \sum_{k<k'} \frac{p(k,j;k',j) - p(k,j)p(k',j)}{(k-1)(k'-1)}. \tag{A28}$$

By Fu (1995, eqns. (29) and (22)), the first and second terms of (A28) are of order $o(n)$ and $O(\frac{\log n}{n})$, 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 A5 (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 (A26) 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 $a_j \sim \mathrm{Poisson}(\frac{\theta}{2}\, \mathbf{E}_n[\tau_j])$ for $1 \leq j \leq b$, as $n \to \infty$.

Proof. By Theorem A4, the probability that nested mutation up to count b occurs tends to 0 as $n \to \infty$. The result then follows from Lemma A4 and Theorem A3. □

It can be checked that exponentially growing popolations clearly satisfy $\sup_{t\geq 0} \lambda(t) < \infty$ and also assumption (A26). The conclusions of Theorems A4 and A5 then hold for the generalized coalescent trees of Griffiths and Tavaré (1998) when $\lambda(t) = e^{\beta t}$ for $t \in \mathbb{R}_+$ for some $\beta > 0$.

Equipped with Theorem A5, 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 $n = \sum_i ia_i$ and its number of latent mutations is $k = \sum_i a_i$. Hence

$$\begin{aligned}
G_{n,k}(x,y) &= \sum_{(a_1,a_2,\ldots)} P(a_1,a_2,\ldots)x^n y^k \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \sum_{(a_1,a_2,\ldots)} x^{\sum ia_i} y^{\sum a_i} \prod_{i\geq 1} \frac{(\theta\bar{\tau}_i/2)^{a_i}}{a_i!} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \sum_{(a_1,a_2,\ldots)} \prod_{i\geq 1} \frac{x^{ia_i} y^{a_i} (\theta\bar{\tau}_i/2)^{a_i}}{a_i!} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \prod_{i\geq 1} \sum_{a_i\geq 0} \frac{x^{ia_i} y^{a_i} (\theta\bar{\tau}_i/2)^{a_i}}{a_i!} \qquad \text{(A29)} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \prod_{i\geq 1} e^{x^i y \theta\bar{\tau}_i/2} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} e^{\frac{\theta}{2} y \sum_i x^i \bar{\tau}_i} \\
&= e^{-\frac{\theta}{2}\sum_i \bar{\tau}_i} \sum_{k=0}^{\infty} \frac{(\frac{\theta}{2})^k y^k}{k!} \left(\sum_i x^i \bar{\tau}_i \right)^k
\end{aligned}$$

as declared in the main text.

A remark on the total number of mutations for large n_1

The stable probabilities observed in the lower three panels of Fig. 6 and in Fig. 4b suggest that the conditional distribution of k_1 given n_1 and a very large n will approach a distribution as n_1 gets larger. That this is the case under constant population size follows from the fact, here as in "A conditional ancestral process for rare variants," that the number of alleles in the Ewens sampling formula is the sum of independent Bernoulli trials (Arratia *et al.* 1992, 2000). The limiting large-n_1 distribution is Poisson, but shifted because there must be at least one mutation to produce $n_1 > 0$ copies, so it is $k_1 - 1$ that is Poisson. See Proposition 3.1 of Yamato (2017).

By the following heuristic argument, we suggest that this result holds more broadly, in particular for growing populations or ones in which $\bar{\tau}_i$ decreases at least as fast with i as in the constant-size model, whatever the reason. In this case, when a mutation occurs it will very likely produce a low-count variant because $\bar{\tau}_i / \sum_i \bar{\tau}_i$ for small i will be much greater than $\bar{\tau}_{n_1} / \sum_i \bar{\tau}_i$ for large n_1. A large-n_1 variant which is due for example to $k_1 = 2$ latent mutations will very likely have a count pattern such as $(a_1 = 1, a_{n_1-1} = 1)$ or $(a_2 = 1, a_{n_1-2} = 1)$ and very unlikely to have one such as $(a_{n_1/2-j} = 1, a_{n_1/2+j} = 1)$ for some small j.

Then we expect the probability (31) of seeing n_1 copies given k_1 latent mutations to be close to

$$p(n_1|k_1; n \text{ large}, \tau) \approx k_1 \frac{\bar{\tau}_{n_1}}{\sum_{i=1}^{n-1} \bar{\tau}_i} \quad \text{when } \mathrm{n}_1 \text{ is large,}$$

because each of the k_1 mutations has a small chance $\bar{\tau}_{n_1} / \sum_i \bar{\tau}_i$ of producing an appropriately large number of copies, the other $k_1 - 1$ mutations being inconsequential to the total count. Multiplying by the Poisson distribution of k_1 in (30) and rearranging gives

$$\begin{aligned}
p(n_1, k_1; n \text{ large}, \tau) &\approx k_1 \frac{\bar{\tau}_{n_1}}{\sum_{i=1}^{n-1} \bar{\tau}_i} \frac{(\frac{\theta\pi_1}{2} \sum_{i=1}^{n-1} \bar{\tau}_i)^{k_1}}{k_1!} e^{-\frac{\theta\pi_1}{2}\sum_{i=1}^{n-1} \bar{\tau}_i} \\
&= \frac{\theta\pi_1}{2} \bar{\tau}_{n_1} \cdot \frac{(\frac{\theta\pi_1}{2} \sum_{i=1}^{n-1} \bar{\tau}_i)^{k_1-1}}{(k_1-1)!} e^{-\frac{\theta\pi_1}{2}\sum_{i=1}^{n-1} \bar{\tau}_i}
\end{aligned}$$

for large n_1. To the left of the · is a probability like (7). To the right of the · is the shifted Poisson distribution, which implicitly averages over the (small) sizes of the $k_1 - 1$ additional mutations.

Editor: G. Coop