The distribution of the number of mutations in the genealogy of a sample from a single population ☆

Yun-Xin Fu[1]

Department of Biostatistics and Data Science, School of Public Health
University of Texas Health Science Center at Houston,
1200 Herman Pressler, Houston, Texas 77030

Abstract

The number K of mutations in the genealogy of a sample of n sequences from a single population is one essential summary statistics in molecular population genetics and is equal to the number of segregating sites in the sample under the infinite-sites model. Although its expectation and variance are the most widely utilized properties, its sampling formula (i.e., probability distribution) is the foundation for all explorations related to K. Recently, it has been established that K is subject to the Central Limit Theorem, and thus has asymptotic normality. However, due to its slow convergence to normality, the finite-sample distribution remains indispensable. Although an analytic sampling formula exists, its numerical application is limited due to susceptibility to error propagation. This paper presents a new sampling formula for K in a random sample of DNA sequences from a neutral locus without recombination, taken from a single population evolving according to the Wright-Fisher model with a constant effective population size, or the constant-in-state model, which allows the effective population size to vary across different coalescent states. The new sampling formula is expressed as the sum of the probabilities of the various ways mutations can manifest in the sample genealogy and achieves simplicity by partitioning mutations into hypothetical atomic clusters that cannot be further divided. Under the Wright-Fisher model with a constant effective population size, the new sampling formula is closely analogous to the celebrated Ewens' sampling formula for the number of distinct alleles in a sample. Numerical computation using the new sampling formula is accurate and is limited only by the burden of enumerating a large number of partitions of a large K. However, significant improvement in efficiency can be achieved by prioritizing the enumeration of partitions with a large number of parts.

Keywords: Probability Distribution, Number of Mutations, Coalescent Theory, Integer Partition, Mathematical Induction, Permutation of Cycle Type

The number K of mutations in the genealogy of a sample of n sequences from a population is one of the most important summary statistics in molecular population genetics. It is also ubiquitous in the analysis of DNA sequence samples, as it can be identified as the number of segregating sites in a sample under the infinite-sites model (Ewens (2004), and also see Fu (2025) for a concise list of theoretical and methodological studies related to K). Therefore, it is essential to characterize its sampling formula (i.e., probability distribution), which is

$$\mathbb{P}_n(m) = \text{Prob}(K = m) \ . \tag{1}$$

Watterson (1975) established that $K = X_2 + ... + X_n$, where X_i represents the number of mutations in the time period of observing i ancestral sequences in a forward process. In the terminology of the Kingman coalescent (Kingman, 1982a,b), X_i is the number of mutations during state i of the coalescent process of the sample, and follows a geometric distribution with parameter $1 - \varphi_i$, where

$$\varphi_i = \frac{\theta}{i - 1 + \theta} \ , \tag{2}$$

where $\theta = 4N\mu$ with N as the effective population size and μ as the mutation rate per sequence per generation. Since X_i is independent X_j $(i \neq j)$, the probability generating function of K is thus the product of those for different X_i, which led Watterson (1975) to derive the expectation and variance of K, and also pointed out that

$$\mathbb{P}_n(0) = \prod_{i=2}^{n} \frac{i-1}{i-1+\theta} \ , \tag{3}$$

which, as well as for $\mathbb{P}_n(m)$ for small m, can also be obtained from direct convolution. Using partial expansion of the probability generating function of K, Tavaré (1984) derived the following sampling formula

$$\mathbb{P}_n(m) = \frac{n-1}{\theta} \sum_{i=2}^{n} (-1)^i \binom{n-2}{i-2} \varphi_i^{m+1} \ . \tag{4}$$

☆Preprint formatted with Elsevier's template. February 4, 2025

[1]Email address: yunxin.fu@uth.tmc.edu

Recently, Fu (2025) established that central limit theory holds for K, meaning that K asymptotically follows a normal distribution. Furthermore, the finite-sample skewness and kurtosis of K are now available, which enhances our understanding of its finite-sample properties. The characterization of K appears to be nearly complete, at least under the Wright-Fisher model with a constant effective population size. However, since K under the infinite-sites model is often considered a counterpart to k, the number of distinct alleles in the sample under the infinite-alleles model, the lack of resemblance of Tavaré's formula (elegant in its own right) to Ewens' sampling formula leaves much to be desired.

For a sample configuration $1^{i_1}....n^{i_n}$ of n sequences, where i_j is the number of distinct alleles with j copies in the sample, it is known (Ewens, 1972; Karlin and McGregor, 1972) that

$$\Pr(1^{i_1}...n^{i_n}) = \frac{n!}{1^{i_1}...n^{i_n} i_1!...i_n!} \frac{\theta^m}{S_n(\theta)} \ , \tag{5}$$

where

$$S_n(\theta) = \theta(1+\theta)...(n-1+\theta) \ . \tag{6}$$

Summing up all sample configurations that have the same number of distinct alleles leads to

$$\Pr(k=m) = \frac{|S_n^m|\theta^m}{S_n(\theta)} \ , \tag{7}$$

where

$$|S_n^m| = \sum_{i_1+...+i_n=m} \frac{n!}{1^{i_1}...n^{i_n} i_1!...i_n!} \ . \tag{8}$$

The fact that K arises in different configurations, and that its probability is the sum of the probabilities of these configurations, is both intuitive and intriguing. This prompts one to wonder whether a similar analogy exists for K . Of course, since K is only asymptotically sufficient for θ (Fu and Li, 1993), any such sampling formulas, if they exist, are expected to be more complex than the Ewens' sampling formula.

It is also unfortunate that Tavaré's sampling formula (4) has a limited scope of applicability in the numerical calculation of $\mathbb{P}_n(m)$, due to error accumulation in the summation of large absolute values with alternating signs. This issue becomes progressively more severe with increasing sample size, often making the formula unusable even for modest sample sizes. Furthermore, it is highly desirable to be able to compute $\mathbb{P}_n(m)$ for a non-constant population.

The primary purpose of this paper is to present and prove a new sampling formula for K, which not only provides a direct analogy to the Ewens' sampling formula, but also serves as a reliable method for computing $\mathbb{P}_n(m)$ under a more general model than the Wright-Fisher model with a constant effective population size.

The main results

The Kingman coalescent (Kingman, 1982a,b) is best known for describing a sample of size n taken from a population that evolves according to the Wright-Fisher model (Ewens, 2004), although the results are robust under several alternative models, including Moran's model. For this reason, we assume the population evolves according to the *constant-in-state* model (Fu, 2025), which is a variation of the Wright-Fisher model. In this model, the effective population size during coalescent state i is constant (N_i), while the assumption of independence among coalescent times remains valid. The constant-in-state model extends the Wright-Fisher model by allowing for populations of non-constant sizes, while still maintaining mathematical tractability. As a result, it has been applied in various contexts (Pybus et al., 2000; Liu and Fu, 2015). Throughout the paper, we restrict our attention to a sample of DNA sequences from a single locus without recombination, where all mutations are selectively neutral.

Integer partitions are essential in this paper and thus warrant a brief introduction. A partition of an integer m refers to expressing m as a sum of positive integers, where the order of the summands does not matter (Riordan, 1958; Abramowitz and Stegun, 1972). For example, $5 = 1+2+2 = 4+1$, so the tuples $(1,2,2)$ and $(4,1)$ are both partitions of 5. The order of parts in a partition is irrelevant, and by convention, the parts are listed in descending order. We will use bold letters to represent integer partitions in this paper and adopt the convention that $\boldsymbol{s} \vdash m$ indicates that $\boldsymbol{s}$ is a partition of m. Thus, $\boldsymbol{s} \vdash m$ represents a tuple of positive integer(s)

$$(k_1, k_2, ...) \tag{9}$$

such that $k_1 \geqslant k_2 \geqslant ...$ and $k_1 + ... = m$. Equivalent but more convenient notation for an integer partition is cycle type (multiplicity form)

$$\boldsymbol{s} = 1^{s_1} 2^{s_2} ... \ , \tag{10}$$

where s_k is the multiplicity (number of occurrences) of integer k in the partition tuple. Therefore, the sample configuration $1^{i_1}...n^{i_n}$ in Eq.(5) is an integer partition of sample size n. For example, $\boldsymbol{s} = (5,2,2,1) \vdash 10$ has cycle type $1^1 2^2 3^0 4^0 5^1$. It is customary to omit the component k^{s_k} in the cycle type when $s_k = 0$, so the above cycle type becomes more concisely $\boldsymbol{s} = 1^1 2^2 5^1$. It is also conventional to consider $1^0 2^0 ...$ as the sole partition of integer 0.

Summation ($\sum$) and multiplication ($\prod$) appear frequently in this paper, often in the context of enumerating integer partitions and counting permutations of cycle type (which will be elaborated on later). We adopt the convention that when there is no summand, $\sum$ is equal to 0, and when there is no factor, $\prod$ is equal to 1.

The main result of this paper is as follows.

Theorem. *Let $\mathbb{P}_n(m)$ denote the probability of having exactly m mutations in the genealogy of a random sample*

of n sequences from a population evolving according to the constant-in-state model. Then,

$$\mathbb{P}_n(m) = \mathbb{P}_n(0) \sum_{\boldsymbol{s} \vdash m} \frac{\prod_{j=1}^{m} \omega_n^{s_j}(j)}{1^{s_1}...m^{s_m} s_1!...s_m!} \ , \tag{11}$$

where the summation is taken over all integer partitions $\boldsymbol{s} = 1^{s_1}2^{s_2}... \vdash m$ and

$$\mathbb{P}_n(0) = \prod_{i=2}^{n} (1 - \varphi_i) \ , \tag{12}$$

$$\omega_n(j) = \sum_{i=2}^{n} \varphi_i^j \ , \tag{13}$$

where

$$\varphi_i = \frac{\theta_i}{i - 1 + \theta_i} \ , \tag{14}$$

where $\theta_i = 4N_i\mu$, with N_i being the effective population size at state i of the coalescent process and μ being the mutation rate per sequence per generation.

Before proceeding with the proof of the Theorem, we introduce some notations for brevity and clarity. Additionally, we present a corollary for the specific case where the population evolves according to the Wright-Fisher model with a constant effective population size.

Define for $\boldsymbol{s} = 1^{s_1}2^{s_2}... \vdash m$

$$c(\boldsymbol{s}) = \frac{m!}{1^{s_1}...m^{s_m} s_1!...s_m!} \ , \tag{15}$$

$$\lambda_n(\boldsymbol{s}) = \prod_{j=1}^{m} \omega_n^{s_j}(j) \ . \tag{16}$$

Then Eq.(11) can be written as

$$\mathbb{P}_n(m) = \frac{\mathbb{P}_n(0)}{m!} \sum_{\boldsymbol{s} \vdash m} c(\boldsymbol{s})\lambda_n(\boldsymbol{s}) \ . \tag{17}$$

It should be noted that $c(\boldsymbol{s})$ defined above is the number of permutations of m different objects conforming to cycle type $\boldsymbol{s}$ (Riordan, 1958; Abramowitz and Stegun, 1972). Its precise meaning in the context of mutations will be elaborated later when we discuss combinatorial interpretation of the Theorem. As far as $c(\boldsymbol{s})$ is concerned, there is no need to specify explicitly the integer to which a partition $\boldsymbol{s}$ belongs, because the integer is equal to

$$\|\boldsymbol{s}\| = \sum_i i \cdot s_i \ . \tag{18}$$

Furthermore it is well known(Abramowitz and Stegun, 1972) that

$$\sum_{(\boldsymbol{s} \vdash n) \wedge (s_1 + s_2 + ... = m)} c(\boldsymbol{s}) = (-1)^{n-m} \begin{bmatrix} n \\ m \end{bmatrix} \ , \tag{19}$$

$$\sum_{\boldsymbol{s} \vdash m} c(\boldsymbol{s}) = m! \ , \tag{20}$$

where $\left[{n \atop m} \right]$ is the Stirling number of the first kind. It follows that $|S_n^m|$ defined by (8) is equal to $(-1)^{n-m}\left[{n \atop m} \right]$, which is known as unsigned Stirling number of the first kind.

Let

$$g_n(m) = \sum_{\boldsymbol{s} \vdash m} c(\boldsymbol{s})\lambda_n(\boldsymbol{s}) \ . \tag{21}$$

Then enumerating over all integer partitions for a given m, it can be verified easily that

$$\begin{aligned}
g_n(0) &= 1, \\
g_n(1) &= \omega_n(1), \\
g_n(2) &= \omega_n(2) + \omega_n^2(1), \\
g_n(3) &= 2\omega_n(3) + 3\omega_n(2)\omega_n(1) + \omega_n^3(1), \\
g_n(4) &= 6\omega_n(4) + 8\omega_n(3)\omega_n(1) + 3\omega_n^2(2) + 6\omega_n(2)\omega_n^2(1) \\
&\quad + \omega_n^4(1), \\
g_n(5) &= 24\omega_n(5) + 30\omega_n(4)\omega_n(1) + 20\omega_n(3)\omega_n(2) \\
&\quad + 20\omega_n(3)\omega_n^2(1) + 15\omega_n^2(2)\omega_n(1) + 10\omega_n(2)\omega_n^3(1) \\
&\quad + \omega_n^5(1) \ ,
\end{aligned} \tag{22}$$

which can be used to compute $\mathbb{P}_n(m)$ quickly for $m \leqslant 5$.

Under the Wright-Fisher model with constant effective size, we have that $\theta = \theta_2 = ... = \theta_n$. It follows from (3) that

$$\mathbb{P}_n(0) = \frac{(n-1)!\theta}{S_n(\theta)} \ . \tag{23}$$

Furthermore all the summands of $\omega_n(j)$ has a common factor θ^j and for a partition $\boldsymbol{s} \vdash m$,

$$\begin{aligned}
\lambda_n(\boldsymbol{s}) &= \prod_{j=1}^{m} \theta^{js_j} \left[\sum_{i=2}^{n} \frac{1}{(i-1+\theta)^j} \right]^{s_j} \\
&= \theta^m \prod_{j=1}^{m} \left[\sum_{i=2}^{n} \frac{1}{(i-1+\theta)^j} \right]^{s_j} \ ,
\end{aligned}$$

which leads to the following corollary

Corollary. *For a random sample of n sequences from a population evolving according to the Wright-Fisher model with constant effective size N, the probability of having exactly m mutations in the sample genealogy is*

$$\mathbb{P}_n(m) = \frac{S'_m(\theta)\theta^m}{S_n(\theta)} \ , \tag{24}$$

where $S_n(\theta)$ is defined by (6) and

$$S'_m(\theta) = \sum_{\boldsymbol{s} \vdash m} c(\boldsymbol{s}) \frac{(n-1)!\theta}{m!} \prod_{j=1}^{m} \left[\sum_{i=2}^{n} \frac{1}{(i-1+\theta)^j} \right]^{s_j} \ ,$$

where the summation is taken over all partitions of m, $c(\boldsymbol{s})$ is defined by (15) and $\theta = 4N\mu$ with μ being the mutation rate per sequence per generation.

We thus have a closely analogous formula to Ewens' sampling formula (7), with two key differences. The first

is that k can take value between 1 and n, while K can take values between 0 and ∞. The second is that, although both $|S_n^i m|$ and $S'_m(\theta)$ are of the form

$$\sum_{\boldsymbol{s}\vdash i} c(\boldsymbol{s}) f(\boldsymbol{s}) \ , \tag{25}$$

$|S_n^m|$ is a summation over partitions of sample size n with m parts where $f(\boldsymbol{s}) = 1$, resulting in the unsigned Stirling number of the second kind (19). In comparison $S'_m(\theta)$ is a summation over all partitions of mutation number m with $f(\boldsymbol{s})$ as a function of θ, partition $\boldsymbol{s}$ and sample size n. This second difference reflects the fact that k is a sufficient summary statistic for θ, so that the detailed configurations leading to k is independent of θ, while K is not a sufficient summary statistic for θ.

Proof of the Theorem

For clarity, let $K_n = K$ for sample size n. Consider first the case of $n = 2$. Since $K_2 = X_2$, which follows a geometric distribution with parameter $1 - \varphi_2$, it follows that

$$\mathbb{P}_2(m) = (1 - \varphi_2)\varphi_2^m \ .$$

Since $\omega_2(j) = \varphi_2^j$, we have for any partition $\boldsymbol{s} = 1^{s_1}2^{s_2}... \vdash m$ that

$$\lambda_2(\boldsymbol{s}) = \prod_{j\geqslant 1} \omega_2^{s_j}(j) = \varphi_2^{s_1\cdot 1+...+s_m\cdot m} = \varphi_2^m \ .$$

Due to Eq.(20) and $p_2(0) = (1 - \varphi_2)$, one arrives at

$$\begin{aligned}\mathbb{P}_2(m) &= \frac{\mathbb{P}_2(0)}{m!}[\sum_{\boldsymbol{s}\vdash m} c(\boldsymbol{s})]\varphi_2^m \\ &= \frac{\mathbb{P}_2(0)}{m!}\sum_{\boldsymbol{s}\vdash m} c(\boldsymbol{s})\lambda_2(\boldsymbol{s}) \ .\end{aligned} \tag{26}$$

Therefore Eq.(17) is true for $n = 2$. Suppose it is true for all the sample sizes up to $n - 1$. Then since $K_n = K_{n-1} + X_n$, the following recurrence relationship holds

$$\mathbb{P}_n(m) = \sum_{i=0}^{m} Pr(X_n = i)Pr(K_{n-1} = m - i) \ . \tag{27}$$

Since $X_n \sim Geom(1 - \varphi_n)$, it follows that

$$\begin{aligned}\mathbb{P}_n(m) &= \sum_{i=0}^{m}(1 - \varphi_n)\varphi_n^i \ \mathbb{P}_{n-1}(m - i) \\ &= \sum_{i=0}^{m} \frac{(1 - \varphi_n)\varphi_n^i \mathbb{P}_{n-1}(0)}{(m - i)!} \sum_{\boldsymbol{s}\vdash(m-i)} c(\boldsymbol{s})\lambda_{n-1}(\boldsymbol{s}) \\ &= \mathbb{P}_n(0)\sum_{i=0}^{m} \sum_{\boldsymbol{s}\vdash(m-i)} \frac{c(\boldsymbol{s})}{(m - i)!}\varphi_n^i \lambda_{n-1}(\boldsymbol{s}) \ .\end{aligned} \tag{28}$$

We shall show that this is in fact an alternative expression of Eq.(11).

From Eq.(22), we have the recurrence relationship

$$\omega_n(i) = \omega_{n-1}(i) + \varphi_n^i$$

which by a standard expansion gives

$$\omega_n^{s_i}(i) = \sum_{j=0}^{s_i} \binom{s_i}{j} (\varphi_n^i)^{s_i-j}\omega_{n-1}^j(i) \ .$$

This immediately leads to

$$\begin{aligned}\lambda_n(\boldsymbol{s}') &= \prod_{j=1}^{m} \omega_n^{s'_j}(j) = \prod_{j=1}^{m}\sum_{s_j=0}^{s'_j} \binom{s'_j}{s_j}(\varphi_n^i)^{s'_j-s_j}\omega_{n-1}^{s_j}(j) \\ &= \sum_{s_1\leqslant s'_1,...,s_m\leqslant s'_m} \prod_{j=1}^{m} \binom{s'_j}{s_j}(\varphi_n^j)^{s'_j-s_j}\omega_{n-1}^{s_j}(j) \\ &= \sum_{s_1\leqslant s'_1,...,s_m\leqslant s'_m} \left[\prod_{j=1}^{m}\binom{s_j}{s_j}\right]\varphi_n^{\sum_j j(s'_j-s_j)} \prod_{j=1}^{m}\omega_{n-1}^{s_j}(j) \ .\end{aligned}$$

For two partitions $\boldsymbol{s} = 1^{s_1}2^{s_2}...$ and $\boldsymbol{s}' = 1^{s'_1}2^{s'_2}...$, define the relationship $\boldsymbol{s} \leqslant \boldsymbol{s}'$ being true if and only if $s_i \leqslant s'_i$, $i = 1, ...$. When $\boldsymbol{s} \leqslant \boldsymbol{s}'$, define their difference partition as

$$\boldsymbol{s}' - \boldsymbol{s} = 1^{s'_1-s_1}2^{s'_2-s_2}... \ .$$

Consequently $\|\boldsymbol{s}'-\boldsymbol{s}\| = \sum_j j(s'_j-s_j)$. Since $\prod_{j=1}^{m}\omega_{n-1}^{s_j}(j) = \lambda_{n-1}(\boldsymbol{s})$, it follows that

$$\lambda_n(\boldsymbol{s}') = \sum_{\boldsymbol{s}\leqslant\boldsymbol{s}'} \left[\prod_{j=1}^{m}\binom{s'_j}{s_j}\right]\varphi_n^{\|\boldsymbol{s}'-\boldsymbol{s}\|}\lambda_{n-1}(\boldsymbol{s}) \ ,$$

where the summation is taken over all partitions $\boldsymbol{s} \leqslant \boldsymbol{s}'$, which include the $\varnothing$. Grouping $\boldsymbol{s}$ according to its $\|\boldsymbol{s}\|$ results in

$$\lambda_n(\boldsymbol{s}') = \sum_{i=0}^{m} \sum_{\boldsymbol{s}\vdash(m-i)\wedge\boldsymbol{s}\leqslant\boldsymbol{s}'} \left[\prod_{j=1}^{m}\binom{s'_j}{s_j}\right]\varphi_n^i\lambda_{n-1}(\boldsymbol{s}) \ .$$

Therefore directly from (17) we have that

$$\begin{aligned}\mathbb{P}_n(m) &= \frac{\mathbb{P}_n(0)}{m!}\sum_{\boldsymbol{s}'\vdash m} c(\boldsymbol{s}')\lambda_n(\boldsymbol{s}') \\ &= \frac{\mathbb{P}_n(0)}{m!}\sum_{\boldsymbol{s}'\vdash m} c(\boldsymbol{s}')\times \\ &\qquad \sum_{i=0}^{m}\sum_{\boldsymbol{s}\vdash(m-i)\wedge\boldsymbol{s}\leqslant\boldsymbol{s}'}\left[\prod_{j=1}^{m}\binom{s'_j}{s_j}\right]\varphi_n^i\lambda_{n-1}(\boldsymbol{s}) \\ &= \mathbb{P}_n(0)\sum_{i=0}^{m}\sum_{\boldsymbol{s}\vdash(m-i)}\varphi_n^i\lambda_{n-1}(\boldsymbol{s})\times \\ &\qquad \sum_{\boldsymbol{s}'\vdash m\wedge\boldsymbol{s}\leqslant\boldsymbol{s}'} \frac{c(\boldsymbol{s}')\prod_{j=1}^{m}\binom{s'_j}{s_j}}{m!} \ .\end{aligned} \tag{29}$$

Since $\boldsymbol{s} \vdash (m-k)$ and $\boldsymbol{s}' \vdash m$, we have

$$\begin{aligned}\frac{c(\boldsymbol{s}')\prod_{j=1}^{m}\binom{s'_j}{s_j}}{m!} &= \frac{\prod_{j=1}^{m}\binom{s'_j}{s_j}}{1^{s'_1}...m^{s'_m}s'_1!...s'_m!} \\ &= \frac{1}{1^{s'_1}...m^{s'_m}\prod_{j=1}^{m}s_j!(s'_j-s_j)!} \\ &= \frac{c(\boldsymbol{s})c(\boldsymbol{s}'-\boldsymbol{s})}{(m-i)!i!} \ .\end{aligned}$$

For a given $\boldsymbol{s} \vdash (m-i)$, enumerating over all $\boldsymbol{s}' \vdash m$ with $\boldsymbol{s}' \geqslant \boldsymbol{s}$ is equivalent to enumerating over all $(\boldsymbol{s}'-\boldsymbol{s}) \vdash i$. Due to Eq.(20), it follows that

$$\begin{aligned}&\sum_{\boldsymbol{s}'\vdash m \wedge \boldsymbol{s}\leqslant\boldsymbol{s}'}\frac{c(\boldsymbol{s}')\prod_{j=1}^{m}\binom{s'_j}{s_j}}{m!} \\ &= \frac{c(\boldsymbol{s})}{(m-i)!i!}\sum_{(\boldsymbol{s}'-\boldsymbol{s})\vdash i}c(\boldsymbol{s}'-\boldsymbol{s}) \\ &= \frac{c(\boldsymbol{s})}{(m-i)!} \ .\end{aligned}$$

Therefore Eq.(29) is the same as Eq.(28), which indicates that Eq.(11) is also true for sample size n. By induction the Theorem is true for all sample sizes. ||

Combinatorial interpretation

Mutations under the infinite-alleles model lead to partitions of the sample, resulting in the probability of a sample partition $1^{i_1}...i_n^{i_n}$ as given in Eq. (5), and the probability of having k distinct alleles as shown in Eq. (7). These probabilities have since been re-derived multiple times and subjected to various combinatorial interpretations (Hoppe, 1984; Donnelly, 1986; Donnelly and Tavaé, 1986; Griffiths and Lessard, 2005).

Under the infinite-alleles model, each distinct allele is unequivocally identified in the sample, but the number of mutations responsible for the observed allelic configuration is ambiguous—only the minimum number of mutations is known. In contrast, under the infinite-sites model, each mutation is unequivocally identified as a segregating site, but the number of mutations alone does not provide sufficient information on how the sequences in a sample are partitioned. This leads to an ambiguous sample configuration, unless more detailed information about each mutation is considered. Therefore, $\mathbb{P}_n(m)$ can only be partitioned according to the configurations of mutations. The theorem suggests that mutations can be partitioned into clusters corresponding to integer partitions, though the meaning of this is not entirely clear. Insight can be gained by examining in detail the cases of two and three mutations.

Consider the case of two mutations first. As for the states at which they may occur, there are only two possibilities. The first is that both mutations occur at the same state, and the second is that they occur at two different states. The probability that there are two mutations in the sample genealogy such that both occur at state i is

$$(1-\varphi_i)\varphi_i^2\prod_{j\neq i}(1-\varphi_j) = \mathbb{P}_n(0)\varphi_i,$$

Therefore the probability that both mutations occur at the same state is $\mathbb{P}_n(0)\sum_i\varphi_i$. Similarly the probability that there are two mutations in the sample genealogy such that one occurs at state i and another at state j is

$$\mathbb{P}_n(0)\varphi_i\varphi_j \ .$$

Enumerating over all $i<j$ (since $j<i$ and $i<j$ represent the same event thus should not be counted twice), leads to that the probability two mutations occur at different state equal to $\mathbb{P}_n(0)\sum_{i<j}\varphi_i\varphi_j$. It follows that

$$\begin{aligned}\mathbb{P}_n(2) &= \mathbb{P}_n(0)\sum_i\varphi_i^2 + \mathbb{P}_n(0)\sum_{i<j}\varphi_i\varphi_j \\ &= \frac{\mathbb{P}_n(0)}{2}\sum_i\varphi_i^2 + \frac{\mathbb{P}_n(0)}{2}[\sum_i\varphi_i^2 + 2\sum_{i<j}\varphi_i\varphi_j] \\ &= \frac{\mathbb{P}_n(0)}{2!}[c(2^1)\lambda_n(2^1) + c(1^2)\lambda_n(1^2)] \ . \qquad (30)\end{aligned}$$

which is Eq.(11) when $m=2$. The split of $\mathbb{P}_n(0)\sum_i\varphi_i^2$ in the second equality into two halves, with the second half combined into the probability of two mutations at different states, is key to arriving at the concise expression in Eq. (30). One interpretation of this partition and reassembly is that the event of observing two mutations at the same state corresponds to the same outcome from two different underlying scenarios: one is that the two mutations arise in a tight cluster within a state, and the other is that the two mutations come from two separated clusters (each containing one mutation) at the same state. Since the sum of the probabilities of these two scenarios is $\mathbb{P}_n(0)\sum_i\varphi_i^2 = \mathbb{P}_n(0)\lambda_n(2^1)$, it is essential to understand how this probability is apportioned between the two scenarios. The derivation of Eq. (30) suggests an even split, which is equivalent to assigning the probability based on the proportion of the number $c(2^1)$ of permutations of cycle type 2^1 among all permutations of two mutations, which is 2!. Therefore, if the two scenarios are separated, the scenario corresponding to a single cluster of two mutations has probability $\frac{\mathbb{P}_n(0)}{2!}c(2^1)\lambda_n(2^1)$. On the other hand, when the scenario of two separated clusters of one mutation each is combined with $\mathbb{P}_n(0)\sum_{i<j}\varphi_i\varphi_j$, it leads to a probability of $\frac{\mathbb{P}_n(0)}{2!}c(1^2)\lambda_n(1^2)$, which represents the scenario where two mutations arise in two clusters, each containing one mutation. These clusters could either be at the same state or at different states.

The above analysis suggests that, although hypothetical and invisible, the concept of a cluster of mutations can help simplify the interpretation of mutations. The cluster of mutations should be the final assignment, as it

will not be further divided or reassigned; otherwise, every mutation will eventually end up in its own cluster. Furthermore, each cluster resides within a single state. We can therefore refer to such clusters as atomic (indivisible) clusters. The rules for apportioning the probability of an event to different partitions of mutations into atomic clusters can be further clarified by examining the case of three mutations.

There are three different observable events for three mutations. The first is that all three occur at a single state, with probability $\mathbb{P}_n(0)\sum_i \varphi_i^3$, the second is that two mutations occur at one state and another at a different state, with probability $\mathbb{P}_n(0)\sum_{i\neq j}\varphi_i^2\varphi_j$, and the third is that three mutations occur at three different states, with probability $\mathbb{P}_n(0)\sum_{i<j<k}\varphi_i\varphi_j\varphi_k$. The assembly of the probabilities leads to

$$\begin{aligned}\mathbb{P}_n(3) &= \mathbb{P}_n(0)\sum_i \varphi_i^3 + \mathbb{P}_n(0)\sum_{i\neq j}\varphi_i^2\varphi_j + \mathbb{P}_n(0)\sum_{i<j<k}\varphi_i\varphi_j\varphi_k \\ &= \frac{\mathbb{P}_n(0)}{6}[2\omega_n(3) + 3\omega_n(2)\omega_n(1) + \omega_n^3(1)] \\ &= \frac{\mathbb{P}_n(0)}{3!}[c(3)\lambda_n(3^1) + c(1^12^1)\lambda_n(1^12^1) + c(1^3)\lambda_n(1^3)] \ .\end{aligned}$$

The first event is now split into three scenarios: one atomic cluster of three mutations, one atomic cluster of two mutations and one atomic cluster of one mutation, and three atomic clusters of one mutation each. The rule for apportioning the total probability $\mathbb{P}_n(0)\sum_i \varphi_i^3$ to the different scenarios is clearly proportional to the number of permutations of cycle types. Similarly, the second event is split into two scenarios: one atomic cluster of two mutations in one state and one cluster in a different state, and two atomic clusters of one mutation each in the same state and one cluster in a different state.

One can continue the analysis for more mutations, but the above analysis suffices the definition of **atomic clusters** as a list of mutations with the following three properties:

1. An atomic cluster cannot be further divided,
2. An atomic cluster resides within a single state,
3. The order (permutation) of mutations within an atomic cluster is cyclic.

A permutation is said to be cyclic if the elements are arranged in circular form so that it only matter what element follows a given element and there is no beginning nor end element. Therefore, a partition of m mutations into a group of atomic clusters corresponds to an integer partition $\boldsymbol{s} = 1^{s_1}2^{s_2}... \vdash m$. The number of permutations of m mutations to confirm with cycle type $\boldsymbol{s}$ can be intuitively derived as follows.

First of all, the number of arrangements of m mutations into s_i of groups (atomic clusters) of size i, $(i = 1, ..., m)$ is given by the standard multinomial coefficient

$$\frac{m!}{(1!)^{s_1}...(m!)^{s_m}}$$

which assumes each group is labeled (identifiable). If groups can be identified only up to their sizes, then the total number of permutations is reduced to

$$\frac{m!}{(1!)^{s_1}...(m!)^{s_m}s_1!...s_m!} \ .$$

If permutation of mutations within each atomic cluster is identifiable, there will be more arrangements. The assumption of permutation within each atomic cluster being cyclic means that the number of permutations for atomic cluster of size i is $(i-1)!$. Therefore the total number of arrangements of m mutations into cycle type $\boldsymbol{s}$ is

$$\frac{m![(1-1)!]^{s_1}...[(m-1)!]^{s_m}}{(1!)^{s_1}...(m!)^{s_m}s_1!...s_m!} = \frac{m!}{1^{s_1}...m^{s_m}s_1!...s_m!} = c(\boldsymbol{s}) \ .$$

We thus arrive at a proposition that can be regarded as a corollary of the main theorem, based on the construct of hypothetical and invisible atomic clusters of mutations

Proposition. *Mutations in the genealogy of a sample of n sequences emerge in atomic clusters. The specification of a list of atomic clusters coincides with an integer partition of the number of mutations m. The probability of mutations emerging in a partition $\boldsymbol{s}$ is given by*

$$\frac{c(\boldsymbol{s})}{m!}\mathbb{P}_n(0)\lambda_n(\boldsymbol{s}) \tag{31}$$

and $\mathbb{P}_n(m)$ is the summation of probabilities of all different partitions of mutations, where $c(\boldsymbol{s}), \mathbb{P}_n(0)$ and $\lambda_n(\boldsymbol{s})$ are defined, respectively, by (15), (12) and (16).

The construct of atomic clusters of mutations and Eq. (31) provide an analogy to Ewens' sampling formula (5). Without the use of hypothetical and invisible boundaries to partition mutations into atomic clusters, attempting to express $\mathbb{P}_n(m)$ as the sum of probabilities for different scenarios of mutation realizations would lead to complicated, multi-layered summations, quickly losing simplicity.

Numerical computation

When numerically evaluating $\mathbb{P}_n(m)$, it is important to verify its accuracy by comparing the results with known properties. In addition to the usual checks for a valid probability distribution, one can compare $\mathbb{P}_n(m)$ against its moments, including

1. **Expectation and variance** (Watterson, 1975; Fu, 2025):

$$\begin{aligned}\mathbb{E}(K) &= \zeta(1) \ , \\ \sigma_K^2 = Var(K) &= \zeta(1) + \zeta(2) \ .\end{aligned}$$

2. **Skewness and kurtonsis** (Fu, 2025):

$$\begin{aligned}\gamma(K) &= [\zeta(1) + 3\zeta(2) + 2\zeta(3)]/(\sigma_K^2)^{3/2} \ , \\ \kappa(K) &= [\zeta(1) + 3\zeta(1)^2 + 7\zeta(2) + 12\zeta(3) \\ &\quad + 6\zeta(2)\zeta(1) + 3\zeta(2)^2 + 6\zeta(4)]/(\sigma_K^2)^2 \ ,\end{aligned}$$

where

$$\zeta(j) = \sum_{i=2}^{n} \left(\frac{\theta_i}{i-1} \right)^j .$$

With these safeguards in place, we evaluated $\mathbb{P}_n(m)$ for various combinations of n,m and θ. In all cases, we encountered no issues with numerical accuracy when using Eq.(11). The impact of sample size n on the computation of $\mathbb{P}_n(m)$ is minor, as it only affects $\omega_n(i)$, which needs to be evaluated once for each different m. However, the computational burden does increase with m, as will be discussed later.

We present two contrasting examples of $\mathbb{P}_n(m)$ here: one corresponding to the Wright-Fisher model with constant effective size, and the other to the constant-in-state model. Figure 1 shows the distribution of $\mathbb{P}_n(m)$ with $\theta = 3$ for three different sample sizes: 50, 500, and 5000. These cases were chosen in part for comparison with the simulated data and the approximating normal distributions from Fu (2025) (Fig. 3, panel b).

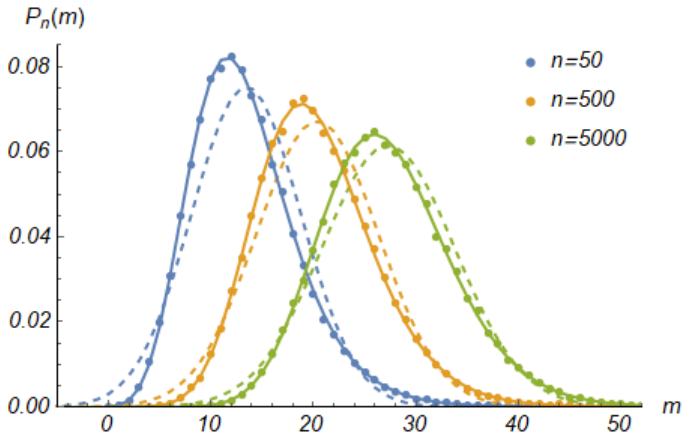

Fig. 1. Probability distribution $\mathbb{P}_n(m)$ with $\theta = 3$ for three samples sizes. For each sample size, solid curve is $\mathbb{P}_n(m)$ computed from (11), dot plot is from coalescent simulation and dashed curve represents the normal approximation. The latter two distributions are from panel (b) of Fig 3 in Fu (2025)

From Fig. 1, it is clear that the exact distribution $\mathbb{P}_n(m)$ for each sample size aligns well with the empirical distribution based on coalescent simulations of 50,000 replicates. This is expected, but without the exact distribution, one might wonder how many replicates are necessary to obtain sufficiently accurate results. With the exact distributions at hand, we can revisit the conclusions of Fu (2025) regarding the asymptotic normality of $\mathbb{P}_n(m)$. One conclusion in Fu (2025) is that the improvement towards normality from $n = 50$ to $n = 500$ is quite noticeable, while the further improvement from $n = 500$ to $n = 5000$ is more subtle. The patterns in Fig. 1 reinforce this conclusion with greater confidence. This is not surprising, given that the speed of convergence of $\mathbb{P}_n(m)$ to normality is approximately $1/\sqrt{ln(n)}$. From $n = 50$ to 500, the change is $1/\sqrt{\ln(500)} - 1/\sqrt{\ln(50)} = 0.515$, while from 500 to 5000 the change is only 0.426. Therefore, after an initial rapid improvement towards normality, further increases in sample size have progressively smaller effects.

For a population evolved according to the constant-in-state model, we considered three scenarios of population dynamics for $n = 100$. The first scenario represents a constant population size, with $\theta_i = 0.407$ for $i = 2, ..., 15$ and 2 for $i = 16, ..., 100$. The second scenario corresponds to a recent population expansion,with $\theta_i = 0.407$ for $i = 2, ..., 15$ and 2 for $i = 16, ..., 100$. The third scenario is a bottleneck followed by expansion, specified by $\theta_i = 0.65$ for $i = 2, 3$, 0.1 for $i = 4, 5, 6$ and 1.425 for $i > 6$. All three scenarios have the same expected number of mutations $E(K) = 5.177$ The mean values of θ for these three scenarios are plotted in Fig. 2 against time into the past, scaled so that 1 unit corresponds to $4N$ generations, where N is the effective population size for the constant population. It is clear that, for the constant population size scenario, the average time to the most recent common ancestor (MRCA) is 1 unit ($4N$ generations). In contrast, for scenarios 2 and 3, the time to the MRCA is shorter due to their smaller effective sizes near the MRCA.

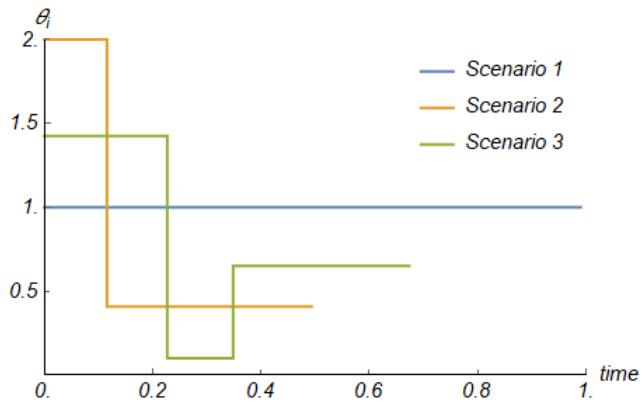

Fig. 2. The mean dynamics of θ_i for a sample of size 100 for three scenarios. Scenario 1: constant model with $\theta = 1$, Scenario 2: $\theta_i = 0.407,\ i = 2, ..., 15$ and $\theta_i = 2,\ i = 16, ..., 100$, and Scenario 3: $\theta_i = 0.65,\ i = 2, 3,\ \theta_i = 0.1,\ i = 4, 5, 6$ and $\theta_i = 1.425,\ i = 7, ..., 100$. All three scenarios lead to $E(K) = 5.177$.

Fig. 3 shows the distributions of $\mathbb{P}_n(m)$ for the three population dynamics scenarios described above. An immediate observation is that these distributions do not differ substantially, despite the seemingly dramatic differences in population size dynamics shown in Fig. 2. While there are noticeable differences around the peak of these distributions, the variations in the tails are minimal. This pattern serves as a reminder that K represents only one

aspect of the genetic polymorphism within a sample. Behind K, the mutations constituting the sample can vary significantly in both type and frequency – a factor that has been leveraged in various data analysis approaches.

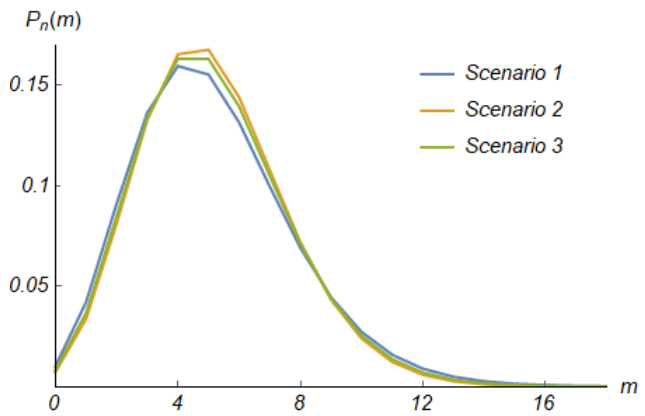

Fig. 3. Probability distribution $\mathbb{P}_n(m)$ computed from (11) for n=100 and three scenarios in Fig. 3

Impact of integer partitions

A key technical issue affecting the computation of $\mathbb{P}_n(m)$ is the number and patterns of integer partitions. Let $p(m)$ denote the number of different partitions of the integer m. The value of $p(m)$ is a critical factor in determining the computational complexity of $\mathbb{P}_n(m)$ using Equation (11), as the computation requires enumerating all possible partitions of m. The function $p(m)$ is a classic problem in combinatorics and number theory, and while a closed-form formula for $p(m)$ is not available, it can be easily computed through enumeration or recurrence relations. For example, the only partition for $m = 0$ is the empty set $\varnothing$, $p(1) = 1$ corresponds to the partition (1), and $p(2) = 2$ corresponds to the partitions (2) and $(1, 1)$. While $p(m)$ increases slowly for small values of m, it grows rapidly as m increases. The thresholds of $10^3, 10^6$ and 10^9 partitions are reached at $m = 22, 61$, and 115, respectively. Table 1 provides values of $p(m)$ for selected values of m up to 150 (Sloane, 2024; Wolfram Research, 2014).

Since every partition contributes positively to the overall probability $\mathbb{P}_n(m)$, there is generally no issue with error propagation when using double-precision computation in modern programming languages and tools. However, for $m > 50$, $p(m)$ increases rapidly, presenting a significant computational challenge. For instance, when $m = 60$, the number of partitions is nearly 1 million, and by the time m reaches 100, the number swells to 190 million. To make matters even more daunting, for $m = 150$, there are over 40 billion partitions! While large numbers may offer elegant solutions in certain contexts, such as in the application of the central limit theorem, they become a major hurdle in the numerical computation of $\mathbb{P}_n(m)$. Therefore, investigating the roles of different types of partitions in the computation of $\mathbb{P}_n(m)$ is important for understanding and improving the efficiency of these calculations.

Table 1. The number $p(m)$ of integer partitions for selected $m \leqslant 150$

m	p(m)	m	p(m)	m	p(m)
0	1	15	175	30	5,604
1	1	16	231	35	14,883
2	2	17	297	40	37,338
3	3	18	385	45	89,135
4	5	19	490	50	204,226
5	7	20	627	55	451,276
6	11	21	792	60	966,467
7	15	22	1,002	65	2,012,558
8	22	23	1,255	70	4,087,968
9	30	24	1,575	75	8,118,264
10	42	25	1,958	80	15,796,476
11	56	26	2,436	90	56,634,173
12	77	27	3,010	100	190,569,292
13	101	28	3,718	125	3,163,127,352
14	135	29	4,565	150	40,853,235,313

Two important characteristics of integer partitions are partition length, which refers to the number of parts in the partition, and maximal size, which denotes the largest value of the parts. In the context of mutation partitions, these correspond to the number of atomic clusters and the maximum number of mutations in a single atomic cluster, respectively. Let $p_i(m)$ represent the number of partitions of length i, and $p'_i(m)$ denote the number of partitions with a maximal size of i. It is known that $p_i(m) = p'_i(m)$ (Slomson, 1991), and thus

$$p(m) = \sum_{i=1}^{m} p_i(m) = \sum_{i=1}^{m} p'_i(m) \ .$$

Then $p_i(m)/p(m)$ is the percentage of partitions of length i among a total $p(m)$ partitions. Similarly $p'_i(m)/p(m)$ is the percentage of partitions with maximal size i among $p(m)$. For a partition $\boldsymbol{s} = 1^{s_1} \ldots$, define $\#(\boldsymbol{s})$ as its number of parts and $\#'(\boldsymbol{s})$ as its maximal size of parts. Then

$$\#(\boldsymbol{s}) = s_1 + s_2 + \ldots \ ,$$
$$\#'(\boldsymbol{s}) = \max\{i : s_i > 0\} \ .$$

Define $\mathbb{P}_n(m, i)$ be the contribution to $\mathbb{P}_n(m)$ by partitions of length i, and similarly $\mathbb{P}'_n(m, i)$ be the contribution to $\mathbb{P}_n(m)$ by partitions $\boldsymbol{s}$ with maximal size i. Then

$$\mathbb{P}_n(m, i) = \frac{\mathbb{P}_n(0)}{m!} \sum_{\boldsymbol{s} \vdash m, \#(\boldsymbol{s}) = i} c(\boldsymbol{s}) \lambda_n(\boldsymbol{s}) \ ,$$

$$\mathbb{P}'_n(m, i) = \frac{\mathbb{P}_n(0)}{m!} \sum_{\boldsymbol{s} \vdash m, \#'(\boldsymbol{s}) = i} c(\boldsymbol{s}) \lambda_n(\boldsymbol{s}) \ ,$$

and $\mathbb{P}_n(m) = \sum_i \mathbb{P}_n(m, i) = \sum_i \mathbb{P}'_n(m, i)$.

Plots of the $p_i(m)/p(m)$ against i/m for three different m with $n = 500$ and $\theta = 10$ are given in Fig. 4. It is clear

that partitions of length in the range $0.2m - 0.3m$ are most abundant. Take the case of $m = 100$ for example (green dotted plot), the peak of $p_i(m)/p(m)$ is around 0.2 for i/m, which means the highest concentration of partitions are those of length 20. Additionally, it is apparent that as m increases, the peak concentration of partitions becomes smaller.

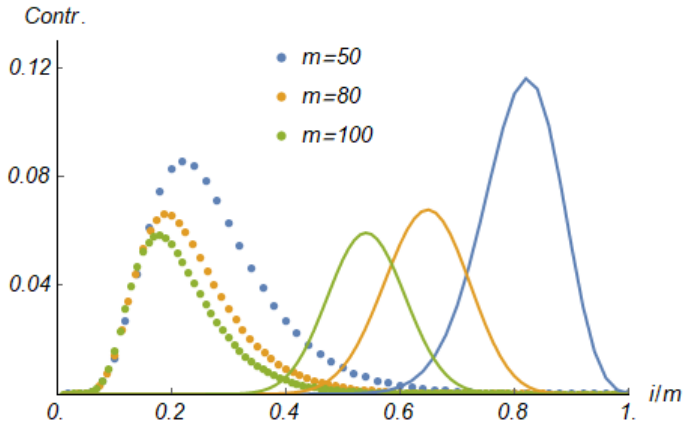

Fig. 4. Percentage of partitions of different lengths (dots) and their contributions(lines) to $\mathbb{P}_n(m)$, for $n = 500$ and $\theta = 10$. Three numbers of mutations considered are $m = 50, 80$ and 100.

Fig. 4 also plots $\mathbb{P}_n(m,i)/\mathbb{P}_n(m)$ against i/m. The contribution towards $\mathbb{P}_n(m)$ is concentrated around 0.55 in the case of $m = 100$ (green curve), indicating that partitions of length around 55 contribute most. It is also clear that the peak of partitions contributing to $\mathbb{P}_n(m)$ gradually shifts to the left with increase of m. Furthermore the density $p_i(n)/p(m)$ and the density $\mathbb{P}_n(m,i)/\mathbb{P}_n(m)$ for a given m overlap only slightly, meaning that in general a small fraction of integer partitions contribute to the overwhelming portion of $\mathbb{P}_n(m)$. In other words, $\mathbb{P}_n(m)$ can typically be computed with sufficient accuracy by evaluating a small portions of partitions. For example, for $m = 100$, 99.99% of the $\mathbb{P}_{500}(100)$ comes from the partitions of length 30 or larger, which account for only 13.4% of the total 190 millions partitions of 100. This feature is quite beneficial, as it makes the challenge of dealing with a large number of partitions a manageable issue.

To test the limits of evaluating $\mathbb{P}_n(m)$ using Eq. (11), we computed $\mathbb{P}_{500}(150)$, which required enumerating over 40 billion partitions of 150. This case was used to examine both the contribution of partitions of different maximal sizes to $\mathbb{P}_n(m)$ and the mean contribution of partitions of various lengths. The results are presented in Fig. 5. From the figure, it is evident that the contribution by partitions of different maximal sizes (green curve) closely overlaps with the percentage of partitions of various maximal sizes (orange curve). This suggests that prioritizing the enumeration of partitions based on their maximal size does not lead to a significant reduction in computational effort.

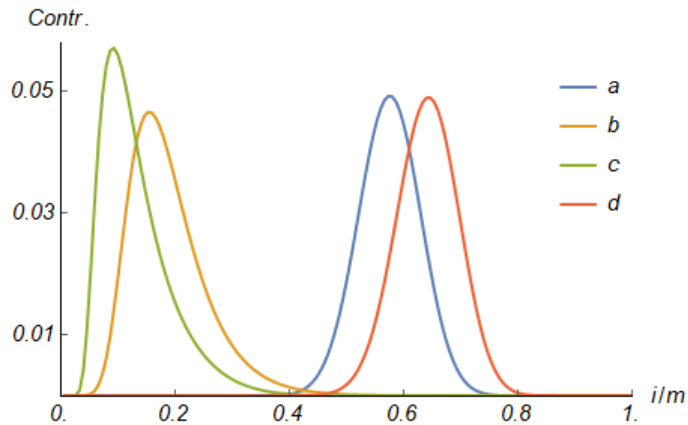

Fig. 5. Contributions of partitions to $\mathbb{P}_{500}(150) = 0.01044$ with $\theta = 20$. (a) Contribution of partitions of different lengths, (b) percentage of $p_i(m)$ in $p(m)$, (c) Contribution by partitions of different maximal sizes and (d) Contribution per partition of different length.

Fig. 5 also shows the per-partition contribution to $\mathbb{P}_n(m)$ for different lengths (red curve). Interestingly, this curve does not coincide with the overall contribution of different lengths (blue curve). In other words, the partition lengths that contribute most to $\mathbb{P}_n(m)$ are not the same as those that contribute the most per partition. However, it is clear that, whether considered collectively or individually, partitions with lengths equal to or larger than $m/2$ contribute most to $\mathbb{P}_n(m)$.

Conditions for the failure of Tavaré formula

It may seem straightforward to compute $\mathbb{P}_n(m)$ using Eq.(4) under the Wright-Fisher model with constant effective size, which is the sum of $n-1$ summands. However, once sample size n is sufficiently large, the binomial coefficient $\binom{n-2}{i-2}$ in the $(i-1)$-th summand can be a very large number, particularly when i is around n/2. As a result, Tavaré formula can often be a summation of some huge summands with alternating sign. To make the matter worse, adjacent summands of opposite signs are typically not close enough in value to cancel each other quickly. When $\mathbb{P}_n(m)$ is about a magnitude 10^{-15} smaller than the absolute values of some summands, error accumulation/propagation is inevitable. This is because modern computers generally carry out floating point calculation in 64-bits format, which leads to 15-17 significant digits, regardless of program languages/tools. There have been improvements in various programming languages, but they are mostly on integer operations and data format. Once floating point operations are performed, the notorious error issue persists.

For illustration, consider $\mathbb{P}_{100}(5)$ with $\theta = 1$. The 4 summands of Eq.(4) for i=48 to 51 are, respectively, $-8.31062 \times 10^{17}, 7.66893 \times 10^{17}, -6.81003 \times 10^{17}, 5.81884 \times 10^{17}$. When adding summands in ascending order of i, the result is 7.089×10^4 in standard double precision Java programming, which leads to $\mathbb{P}_n(5) = 7.01 \times 10^6$, while the correct value is $\mathbb{P}_{100}(5) = 0.1553$ according to Eq. (11). One can explore different orders of adding summands and the final result does change slightly, but all results in qualitatively wrong answer.

In our attempt to improve the applicability of Eq.(4), we explored three different implementations, which were standard programming in Java (version 22) using double precision, Java programming using BigDecimal class which has unlimited precision for integer computation and improved precision for decimal numbers, and symbolic programming by Mathematica V10 (Wolfram Research, 2014), which is supposed to have unlimited precision in applicable numerical computations. We observed that significant error usually started with smaller m, as expected, because the power factor φ_i^{m+1} with large m can help to lower the absolute value of a summand, thus reducing the error accumulation. Also expected is that for a given combination of m and θ, errors become more pronounced with increasing sample size. This prompted us to exam the threshold sample size that leads to significant error in $\mathbb{P}_n(m)$. Table 2 gives the these sample sizes for several values of θ in $\mathbb{P}_n(1)$ and $\mathbb{P}_n(5)$. We started with $m = 1$ instead of $m = 0$ since $\mathbb{P}_n(0)$ can be computed directly by Eq.(3) without using Eq.(4). We define failure of Eq.(4) as a situation in which $\mathbb{P}_n(m)$ has at least 10% error when compared with $\mathbb{P}_n(m)$ from Eq.(11).

Table 2. Failure conditions for computing $\mathbb{P}_n(m)$ using Eq.(4) by three different programming methods

θ	Java double	BigDecimal	Mathematica
0.05	57^a,80^b	60,87	62,82
0.10	58,80	59,83	58,83
0.50	52,75	58,80	58,81
1.00	48,70	55,75	54,76
5.00	38,52	39,55	40,55
10.0	27,38	30,45	30,41

Note. Error rate= $100 \times [\mathbb{P}'_n(m) - \mathbb{P}_n(m)]/\mathbb{P}_n(m)$ where $\mathbb{P}_n(m)$ is computed by Eq.(11) and $\mathbb{P}'_n(m)$ by Eq.(4). ^{a} sample size at or larger will lead to $\mathbb{P}'_n(1)$ with at least 10% error, ^{b} sample size at or larger will lead to $\mathbb{P}'_n(5)$ with at least 10% error.

It is clear from Table 2 that for a wide range of θ, Eq.(4) starts to fail for $\mathbb{P}_n(1)$ for sample size around $30 - 60$, and starts to fail for $\mathbb{P}_n(5)$ for sample sizes around $40 - 85$. Higher precision programmings push the threshold sample sizes only slightly higher, which validates the early statement that internal computation in modern computers is mostly 64-bits format, regardless of programming languages/environments. Consequently Tavaré formula is not recommended in general for numerical computation when $n > 50$, particularly when the entire range of m need to be considered.

Summary

A new formula is derived in this paper for the probability of the number K of mutations in the genealogy of a random sample of DNA sequences from a single locus without recombination, drawn from a population evolved according to the constant-in-state model. In the case of constant effective population size, this formula closely resembles Ewens' sampling formula for the number of distinct alleles in the sample. The new sampling formula is based on the probabilities of different partitions of mutations into coalescent states and achieves significant simplification by recognizing and partitioning mutations into atomic clusters – which are indivisible, bounded within coalescent states, and where permutation of mutations is cyclic. Without the concept of atomic clusters, expressing $\mathbb{P}_n(m)$ as the sum of probabilities for different realizations of mutations in the sample genealogy would lead to complex, multi-layered summations whose complexity grows rapidly with the number of mutations.

The new formula can be used effectively for numerical computation with minimal concern for accuracy, since every summand contributes positively. While the efficiency of the computation remains largely unaffected by sample size, it is impacted by the number of mutations due to the need to enumerate all possible partitions of the given number of mutations. In typical applications, the formula should handle computations efficiently with little or no optimization of the algorithm. Moreover, partitions with a larger number of atomic clusters contribute more to the overall probability, so computation efficiency can be improved by prioritizing the evaluation of partitions based on the number of atomic clusters.

In contrast, Tavaré's formula has a limited scope of applicability, primarily due to the error-prone summation of large summands with alternating signs. As such, it is not recommended for general numerical computation of the distribution of K, especially when the sample size is 50 or greater. However, it may still be useful for computing $\mathbb{P}_n(m)$ for large m with smaller sample sizes, potentially in conjunction with the new formula. A Java package implementing the new formula for $\mathbb{P}_n(m)$ is available from the author upon request.

The relevance of permutations of cycle type and integer partitions to population genetics was first highlighted through Ewens' sampling formulas under the infinite-alleles model (Ewens, 1972; Karlin and McGregor, 1972), but has since had limited application to other summary statistics. The newly derived sampling formula also successfully connects to permutations of cycle type and integer partitions. The links between theoretical population genetics and the fields of combinatorics and number theory may be deeper

than previously recognized. Consequently more summary statistics related to K can be expected to possess such characteristics.

References

Abramowitz, M., Stegun, I.A., 1972. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. 9 ed., Dover Publications, Inc., New York.

Donnelly, P., 1986. Partition structures, polya urns, the Ewens' sampling formula and the age of alleles. Theor. Popul. Biol. 30, 271–288.

Donnelly, P., Tavaé, S., 1986. The ages of alleles and a coalescent. Adv. Appl. Probab. 18, 1–19.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3, 87–112.

Ewens, W.J., 2004. Mathematical Population Genetics. Springer-Verlag, New York.

Fu, Y.X., 2025. The central limit theorem for the number of mutations in the genealogy of a sample from a large population. Theor. Popul. Biol. , (in press) (bioRxiv doi: https://doi.org/10.1101/2025.01.23.634620).

Fu, Y.X., Li, W.H., 1993. Maximum likelihood estimation of population parameters. Genetics 134, 1261–1270.

Griffiths, R.C., Lessard, S., 2005. Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to the ages of alleles. Theor. Popul. Biol. 68, 167–177s.

Hoppe, F.M., 1984. Polya-like urns and the Ewens' sampling formula. J. Math. Biol. 20, 91–94.

Karlin, S., McGregor, J.L., 1972. Addendum to a paper of W. Ewens. Theor. Popul. Biol. 3, 113–116.

Kingman, J.F.C., 1982a. The coalescent. Stochastic Processes and Their Applications 13, 235–248.

Kingman, J.F.C., 1982b. On the genealogy of large populations. J. Appl. Probab. 19A, 27–43.

Liu, X., Fu, Y.X., 2015. Exploring population size changes using snp frequency spectra. Nat Genet 47, 555–559.

Pybus, O., Rambaut, A., Harvey, P., 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155, 1429—1437.

Riordan, J., 1958. Introduction to combinatorial analysis. John Wiley, New York.

Sloane, N.J.A., 2024. Seauence A000041: The on-line encyclopedia of integer sequences: https://oeis.org/ .

Slomson, A., 1991. An Introduction to Combinatrics. Chapman and Hall Mathematics, London.

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

Watterson, G.A., 1975. On the number of segregation sites. Theor. Popul. Biol. 7, 256–276.

Wolfram Research, I., 2014. Mathematica V10. Wolfram Research, Inc., Champaign, Illinois.