Theoretical Population Biology 121 (2018) 60–71

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Single and simultaneous binary mergers in Wright-Fisher genealogies

Andrew Melfi, Divakar Viswanath *

Department of Mathematics, University of Michigan, United States

ARTICLE INFO

Article history:
Received 13 October 2017
Available online 12 April 2018

Keywords:
Wright–Fisher model
Kingman coalescent
Large samples
Convergence theory

ABSTRACT

The Kingman coalescent is a commonly used model in genetics, which is often justified with reference to the Wright-Fisher (WF) model. Current proofs of convergence of WF and other models to the Kingman coalescent assume a constant sample size. However, sample sizes have become quite large in human genetics. Therefore, we develop a convergence theory that allows the sample size to increase with population size. If the haploid population size is N and the sample size is $N^{1/3-\epsilon}$, $\epsilon > 0$, we prove that Wright-Fisher genealogies involve at most a single binary merger in each generation with probability converging to 1 in the limit of large N. Single binary merger or no merger in each generation of the genealogy implies that the Kingman partition distribution is obtained exactly. If the sample size is $N^{1/2-\epsilon}$, Wright-Fisher genealogies may involve simultaneous binary mergers in a single generation but do not involve triple mergers in the large N limit. The asymptotic theory is verified using numerical calculations. Variable population sizes are handled algorithmically. It is found that even distant bottlenecks can increase the probability of triple mergers as well as simultaneous binary mergers in WF genealogies.

1. Introduction

The Kingman coalescent (Kingman, 1982a, b) is a mathematical model of the genealogy of n haploid samples. If k lineages are present in some earlier generation, those lineages induce a partition of the n current samples into k. For convenience, we will refer to lineages present in earlier generations as ancestral samples.[1]

One of Kingman's motivations in deriving the coalescent (Kingman, 1982a, b, 2000) was to gain an understanding of the structure of Ewens' sampling formula (Ewens, 1972; Durrett, 2008). The coalescent gives an almost instantaneous derivation of Ewen's sampling formula, and Ewens' sampling formula is exact under the coalescent approximation. The coalescent is perfectly memoryless in the following sense: at every coalescence exactly two ancestral samples are picked at random (without regard to the number or inter-relationship of their descendants) and deemed to have a common parent. That memoryless property is the chief reason for its simplicity and usefulness.

The Wright-Fisher (WF) model says that if a haploid population of size N_1 produces N_2 children in the next generation, the split of the N_2 children between N_1 parents is multinomial (Durrett, 2008). In the backward in time genealogical process, the k samples in a generation choose parents from their parental generation independently, with each individual of the parental generation being equally likely to be chosen. The individuals of the parental generation that turn out to be parents of any of the k samples constitute the parental sample. Such a passage from a sample to its parental sample will be referred to as a backward WF step. The WF genealogy of a sample is a sequence of backward WF steps until an ancestral generation with a single ancestral sample is reached.

The WF model assumes non-overlapping generations, and there is no attempt to model pedigree relationships in WF (Wakeley et al., 2012). Genealogies in WF as well as other exchangeable models have been proven to converge to the Kingman coalescent (Kingman, 1982b; Möhle, 2000; Möhle and Sagitov, 2001). These proofs assume the sample size to be fixed and constant with $N \to \infty$, where N is the population size. Rapid progress in human genetics has led to sample sizes that are greater than the baseline assumption of an effective population size of N with $N = 2 \times 10^4$ (Karczewski et al., 2016; Sudlow et al., 2015). Thus, there is a need to advance convergence theory beyond the assumption of constant sample size. The beginning of such a convergence theory is presented in this paper by considering the genealogical coalescence process using Kingman's model as well as the WF model.

Approximation of a single WF generation using the coalescent. If the sample size n is constant, $N \to \infty$, and N generations of WF are identified with a single unit of time in the Kingman coalescent, WF genealogies converge to the Kingman coalescent (Kingman, 1982a, b). For constant sample size n and large N, any mergers in a single WF generation are single binary mergers with probability converging to 1. However, if the sample size n is

* Corresponding author.
E-mail addresses: melfi@umich.edu (A. Melfi), divakar@umich.edu (D. Viswanath).

[1] The "ancestral sample" nomenclature is more intuitive for our purposes. However, in the context of the coalescent, the same concept is referred to as "lineage" or "ancestral lineage" (Griffiths, 2006; Griffiths and Tavaré, 1998; Tavaré, 1984).

https://doi.org/10.1016/j.tpb.2018.04.001
0040-5809/

comparable to N, there will be simultaneous binary mergers as well as triple mergers in a single WF generation (Aldous, 1989; Fu, 2006). A single WF generation corresponds to a time interval of $1/N$ in the Kingman coalescent. Because the Kingman coalescent employs a continuous time Poisson process and sets the rate of binary mergers equal to $\frac{n(n-1)}{2N}$, it may still be able to capture the multiple mergers that occur in a single WF generation (Bhaskar et al., 2014; Fu, 2006).

Nevertheless, the coalescent and WF will not produce identically distributed genealogies. There are two differences, and the first difference lies in differing rates of coalescence. The rate at which lineages disappear in a single generation is approximately a function of n/N for both WF and Kingman but it is not the same function (Fu, 2006, Fig. 3). However, the disparity between rates can be mostly eliminated by making the population size N in the Kingman coalescent an appropriate function of the sample size n. In particular, suppose there are n samples in a WF generation with parental population size equal to N. In Kingman, the parental population size can be taken to be N' with

$$s_{WF}\left(\frac{n}{N}\right) = s_K\left(\frac{n}{N'}\right), \tag{1}$$

where s_{WF} and s_K are functions depicted in Fig. 3 of Fu (2006).

Another difference between WF and the Kingman model for large sample sizes n lies in generating partitions whose probability distributions are different. This difference is noteworthy because there is no obvious way to eliminate it. Suppose 10 samples in a single generation are known to have one of two parents from the previous generation, with both parents known to have at least one child among the 10 samples. Under WF, the split of the 10 unlabeled children between the two labeled parents is binomial. That means that $1+9$, $5+5$, and $9+1$ splits have probabilities equal to

$$\frac{\binom{10}{1}}{2^{10}-2} = 1\%,\quad \frac{\binom{10}{5}}{2^{10}-2} = 25\%,\quad \frac{\binom{10}{9}}{2^{10}-2} = 1\%,$$

respectively. If a single generation of WF is modeled using Kingman, the splits under the same condition would all have probability equal to 1/9 (Durrett, 2008, page 13, Theorem 1.6) (Griffiths and Tavaré, 1998). Thus, it is clear that although the Kingman coalescent can produce simultaneous binary mergers as well as multiple mergers over a time interval corresponding to a single generation, the partitions it produces will have a different distribution from that of WF.

Convergence theory for sample sizes that increase with N. As implied by the classic birthday problem and its variants (Aldous, 1989), some two individuals in a sample of size $N^{1/2}$, assuming a fixed population size of N, will have a common parent (binary merger) with a probability of $1-e^{-1/2}$ in the limit of large N. In samples of size $N^{1/2-\epsilon}$, $\epsilon > 0$, there are no common parents in a typical generation in the limit of large N, and when there are common parents, it is reasonable to assume that at most two individuals have a common parent. However, when the sample size is $N^{2/3}$, some three samples will have a common parent (triple merger) with a probability of $1-e^{-1/6}$ in the limit of large N. For sample sizes in-between $N^{1/2}$ and $N^{2/3}$, there will be simultaneous binary mergers (between distinct pairs of samples) in a single generation with high probability. By our convention, quadruple or higher mergers also count as triple mergers.

In the Kingman coalescent, every coalescence is a single binary merger. If the sample size is $N^{1/3-\epsilon}$, $\epsilon > 0$, we prove that each backward WF step involves at most a single binary merger with probability converging to 1 in the limit of large N. Thus, for such sample sizes, the distribution of partitions (with each part in the partition being the subset of current samples descended from an ancestral sample) will converge to the Kingman partition distribution.

It has been suggested that simultaneous binary mergers may cause less divergence from summary statistics such as the sample frequency spectrum than triple and higher multiple mergers (Bhaskar et al., 2014; Davies et al., 2007). We prove a result (Corollary 2) that may partially support that suggestion. In addition, we prove that WF genealogies do not involve triple mergers for sample sizes of $N^{1/2-\epsilon}$. In fact, our results are more detailed. For example, we prove that for sample sizes of $N^{2/5-\epsilon}$ each backward WF step in the genealogy has either zero, one, or two binary mergers with probability converging to 1 for large N. That result is in turn extended to allow c or fewer binary mergers with $c = 3, 4, \ldots$.

We develop algorithms to compute the probability that the genealogy of a sample involves at most a single binary merger in each backward WF step and the probability that there are no triple mergers. Numerical computations using these algorithms show that the asymptotic theory applies to even $N = 10^3$.

The algorithms can handle demographic histories with varying population sizes. Thus, we are able to apply the algorithms to different models of human demography. It is found that even distant bottlenecks can increase the likelihood of WF genealogies with simultaneous binary mergers or triple mergers. A Python/C implementation of the algorithms we derive is posted at `github.com/melfiand/lsample`.

Convergence of the WF sample frequency spectrum. Suppose a sample of size n is polymorphic at a certain nucleotide location. Under the Kingman model and in the limit of zero mutation rate, the probability that k out of n samples are mutants is equal to

$$\frac{1/k}{1+\frac{1}{2}+\cdots+\frac{1}{n-1}}$$

for $k = 1, \ldots, n-1$ (Durrett, 2008; Griffiths and Tavaré, 1998). We prove that the WF sample frequency spectrum converges to the same distribution for samples of size $N^{1/3-\epsilon}$ or smaller in the limit of large N and zero mutation rate.

The $N^{1/3}$ cut-off is almost certainly too pessimistic. A summary statistic such as the sample frequency spectrum partitions the sample into only two sets – samples which have been hit with a mutation and samples which have not been hit with a mutation – under the assumption that the probability of two mutations in the genealogical tree is negligible. In contrast, convergence to the Kingman partition distribution requires partition distributions to match at every level of the genealogical tree. Our proof of convergence assumes all mergers to be single binary mergers and therefore relies on convergence to the Kingman partition distribution as an intermediate step.

The sample frequency spectrum is used in demographic inference and other applications (Gravel et al., 2011; Keinan et al., 2007; Tennessen et al., 2012). Because of its pertinence to applications, the departure of the WF sample frequency spectrum from that of the coalescent has attracted attention. Wakeley and Takahashi (2003) observed (relying on the earlier work of Fisher) that if the sample size is $n = Nx$, where N is the parental population size, the number of parents after a single backward WF step is equal to $N(1-e^{-x})$ in expectation (with a standard deviation that is proportional to $\sqrt{N}$). If $2N\tau_1(x)$ is the size of the external branches in the genealogical tree (in our terminology, the external branch size is equal to the sum of the number of current samples and the number of ancestral samples with exactly one descendant), Wakeley and Takahashi (2003) derived a recurrence for $\tau_1(x)$. From that recurrence, they deduced that the probability of a single mutant in a population sized sample (for which $n = N$) exceeds its Kingman value by 12.05% in the limit of large N. The departure from the Kingman value for the probability of k mutants decreases rapidly with k. These results have been confirmed by Bhaskar et al. (2014).

Fu (2006) derived an exact coalescent for WF. Like Wakeley and Takahashi (2003), he found the main effect of large sample sizes

on the sample frequency spectrum of WF relative to the coalescent to be due to greater external branch lengths. He also showed the Kingman coalescent to be faster than WF for large samples, while noting that simultaneous binary mergers were dominant even for sample sizes large enough to cause triple mergers with appreciable probability.

Whereas Fu (2006) used computer simulations of the exact WF coalescent to study the sample frequency spectrum, Bhaskar et al. (2014) derived exact recurrences for the sample frequency spectrum as well as the expected number of triple mergers and other genealogical quantities. The algorithms of Bhaskar et al. (2014) are applicable to demographic histories with varying population sizes. Rapid population expansion as well as large sample effects increase the probability of single mutants.

In part of the literature on large samples, the focus is on rates of coalescence and the number of ancestral samples as a function of the ancestral generation, with the Kingman model assumed. Tavaré (1984) obtained formulas for the size of the ancestral sample (number of lineages) as a function of the ancestral generation, assuming fixed population size. Griffiths and Tavaré (1998) obtained formulas that allowed the population size to vary. These formulas employ a sum whose terms alternate in sign and are inaccurate when the sample size is large, even assuming the coalescent approximation. Thus, Griffiths (2006) obtained asymptotic approximations that are better numerically for large samples. Other authors (Chen and Chen, 2013; Polanski and Kimmel, 2003; Polanski et al., 2017) have extended this work to handle coalescence and inter-coalescence times. In particular, Chen et al. (2015) have observed that the number of segregating sites, an important statistic introduced by Watterson (1975) and which marked the shift from infinite alleles to the infinite sites model (Durrett, 2008), appears to be more robust under the coalescent approximation than the sample frequency spectrum for large sample sizes. With regard to the sample frequency spectrum, the difficulties due to alternating signs can be handled using a recurrence of Tavaré (1984) as shown by Bhaskar et al. (2014).

2. Convergence theory for sample sizes that increase with N

The coalescent consists of two independent stochastic processes (Kingman, 1982b). Let $[n]$ denote the set $\{1, 2, \ldots, n\}$, which is the current sample. A partition of the set $[n]$ is a set of nonempty subsets of $[n]$ that are pairwise disjoint and whose union is the set $[n]$. In Kingman's coalescent, the partition $\{A_1, \ldots, A_k\}$ of $[n]$ is initialized to $\{\{1\}, \ldots, \{n\}\}$ with $k = n$. At each step, two sets A_i and A_j are chosen, with each of the possible $k(k-1)/2$ choices equally likely, and the two sets are replaced by their union $A_i \cup A_j$. This stochastic process, which governs the evolution of partitions of $[n]$, has been called the jump chain (Kingman, 1982b). A partition of $[n]$ with k parts signifies an ancestral sample (in some earlier generation) of size k, with each ancestral sample denoted by the set of its descendants in the current sample. The merging of two partitions corresponds to two ancestral samples having a common parent so that the number of ancestral samples is reduced by 1.

The other part of the coalescent is the so-called death process (Kingman, 1982b), which governs the timing of the coalescence events. The death process is a continuous time Poisson process of varying rate, with the rate being $k(k-1)/2$ when the number of ancestral samples is k. The connection to the WF model is made by equating a unit of time in the death process with N WF generations.

The jump chain and the death process are independent, and the death chain does not play any role in the convergence to the Kingman partition distribution. The death process governs the rates of coalescence, which can be adjusted independently, as shown in (1).

The following theorem of Kingman (1982b) characterizes the jump chain completely via the Kingman partition distribution and does not depend upon the death chain:

Theorem 1. *Suppose that the coalescent is run until the partition of $[n]$ consists of exactly k sets. If $|A_j| = n_j$ is the cardinality of A_j, the probability that the partition into k sets is $\{A_1, \ldots, A_k\}$ is equal to*

$$\frac{(n-k)!k!(k-1)!}{n!(n-1)!} n_1! n_2! \ldots n_k!.$$

All theorems and corollaries stated in this section are proved in the Appendix. The numbering in this section is the same as in the Appendix. For the above theorem, the Appendix gives a combinatorial proof of the Kingman partition distribution in the spirit of Griffiths and Lessard (2005). Kingman's proof is recursive (Kingman, 1982b; Durrett, 2008).

Simultaneous binary mergers in backward WF steps may cause less deviation because they can be produced by the coalescent with appreciable probability, as shown by the following corollary:

Corollary 2. *Suppose the set $\{\{1\}, \ldots, \{n\}\}$ undergoes k coalescences resulting in a partition into $n-k$ sets. The probability $q(k, n)$ that each set in the resulting partition is of size 1 or 2 is given by $q(k, n) = \frac{(n-k)^{\underline{k}}}{(n-1)^{\underline{k}}}$. If $3k \leq n$ and $k \geq 2$, we have $\exp\left(-\frac{k^2}{2n}\right) \geq q(k, n) \geq \exp\left(-\frac{7k^2}{n}\right) \geq 1 - \frac{7k^2}{n}$.*

In this corollary, the falling power $n(n-1)\ldots(n-k+1)$ is denoted $n^{\underline{k}}$ as recommended by Knuth (Graham et al., 1994; Knuth, 1997). The corollary implies that k simultaneous binary mergers are produced with probability close to 1 as a result of k steps of the jump chain if k is much less than $\sqrt{n}$, where n is the sample size. Therefore, we will look at bounds on n in terms of the population size N that allow only single binary mergers in the WF genealogy of the sample (with high probability) as well as bounds that allow simultaneous binary mergers.

For a constant population size equal to N, the following theorem gives sample sizes that ensure that each backward WF step in the genealogy has at most a single binary merger:

Theorem 4. *The WF genealogy of a sample of size $N^{1/3-\epsilon}$, $\epsilon > 0$, involves at most a single binary merger per generation with probability converging to 1 in the limit of large N.*

This theorem does not consider rates of coalescence. The theorem only claims that the probability that there are either simultaneous binary mergers or triple mergers in the WF genealogy of the sample goes to zero for large N for sample sizes smaller than $N^{1/3-\epsilon}$. However, for such sample sizes, the rates of mergers in WF genealogies agree with the rates of the coalescent (the death process) asymptotically, as will become clear from the statement and proof of a theorem about the sample frequency spectrum given later.

In light of Theorem 1, suppose we look for a bound on the sample size that ensures that every backward WF step consists of either zero, one, or two binary mergers. We then have the following theorem:

Theorem 9. *Each backward WF step in the genealogy of a sample of size $N^{2/5-\epsilon}$, $\epsilon > 0$, has zero, one, or two binary mergers with probability converging to 1 for large N.*

For another interpretation of this theorem, we may define the mod-2 coalescent in analogy with the Kingman coalescent. In an ancestral sample of size k, the mod-2 coalescent picks 4 individuals at random, divides them into two pairs, and merges both pairs. The merger can be thought of as a union of sets, with each set being the set of descendants present in the current sample of an individual in the ancestral sample. It is equivalent to ancestral individuals in both pairs finding common parents, the parents of the two pairs being distinct. The above theorem may then be interpreted as

saying that the WF coalescent of samples of size $N^{2/5-\epsilon}$ or less is a mixture of the coalescent and the mod-2 coalescent, with the proportion of the mixture varying with sample size.

More generally, we may allow c simultaneous binary mergers rather than just 2. We have the following theorem:

Theorem 9 (*General Case*). *The probability that each backward WF step in the genealogy of a sample of size $N^{\frac{c}{2c+1}-\epsilon}$, $\epsilon > 0$, consists only of binary mergers, with the number of binary mergers in any generation being one of $0, 1, \ldots, c$, converges to 1 in the limit of large N.*

It is clear from this theorem that triple mergers may occur for sample sizes of the order $N^{1/2}$ or higher. If N is large and the sample size is smaller than $N^{1/2-\epsilon}$, it follows that all mergers in backward WF steps are simultaneous binary mergers.

Let $f(k, n)$ be the probability that k out of n samples are mutants conditional on exactly one mutation in the genealogy of the sample. Let $\mathcal{H}_n$ denote the harmonic number $1 + \frac{1}{2} + \cdots + \frac{1}{n}$. The coalescent implies $\tilde{f}(k, n) = \frac{1/k}{\mathcal{H}_{n-1}}$ for $k = 1, \ldots, n-1$ in the limit of zero mutation rate as noted in the introduction. The following theorem shows that the WF sample frequency spectrum converges to that of the coalescent for sample sizes smaller than $N^{1/3-\epsilon}$:

Theorem 13. *Let $f_{WF}(k, n)$ be the probability that k out of n samples are mutants conditional on exactly one mutation in the WF genealogy of the sample. Then the total variation distance*

$$\frac{1}{2}\sum_{k=1}^{n-1}\left|f_{WF}(k, n) - \frac{1/k}{\mathcal{H}_{n-1}}\right| \to 0$$

for $n \leq N^{1/3-\epsilon}$, $\epsilon > 0$, in the limit of zero mutation and large N.

3. Algorithms for varying population sizes

For any sample size $n > 2$ and finite N, the probability that the WF genealogy of the sample includes simultaneous binary mergers or triple mergers is strictly greater than zero. Indeed, the probability of such events in a single backward WF step is strictly greater than zero. However, by Theorem 4, the probability that the WF genealogy includes only single binary mergers converges to 1 in the limit $N \to \infty$ if $n \leq N^{1/3-\epsilon}$, where N is the constant population size.

In this section, we derive an algorithm that calculates the probability that the WF genealogy involves only single binary mergers. We derive another algorithm that calculates the probability that the genealogy of a sample of size n does not involve even a single triple merger. Both algorithms allow variable population sizes and may also be used to verify some of the asymptotic results.

Let $p(0, n, N)$ be the probability that a sample of size n does not undergo any merger in a single Wright-Fisher step. Then

$$p(0, n, N) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{n-1}{N}\right),$$

with N the population size of the parental generation. Let $p(k, n, N)$ be the probability of exactly k binary mergers and no triple mergers in a backward WF step with parental population size equal to $2N$. Then

$$p(k, n, N) = \binom{n}{2k}(2k-1)(2k-3)\ldots 3.1\left(\frac{1}{N}\right)^k \times \left(1 - \frac{1}{N}\right)\cdots\left(1 - \frac{n-k-1}{N}\right)$$

for $0 \leq 2k \leq n$. The formula is valid for $n > N$. The formula may be justified as follows. First, we may choose $2k$ samples to participate in k simultaneous binary mergers in $\binom{n}{2k}$ ways. To group the $2k$ samples into k pairs, the first of the chosen samples may be paired in $(2k-1)$ ways, the first of the remaining $2k-2$ samples may be paired in $2k-3$ ways, and so on. Thus, the total number of pairings is $(2k-1)(2k-3)\ldots 3.1$. For each pair, the probability that the two samples in the pair have a common parent is $\frac{1}{N}$. The remaining factors in the formula give the probability that the k pairs as well as the remaining $n-2k$ samples have $n-k$ distinct parents.

Probability of at most a single binary merger in any generation

For the current generation from which a sample of n is taken, we assume $t = 0$. Let $N(t)$ be the haploid population size t ancestral generations ago. To calculate the probability that the WF genealogy of the sample has at most a single binary merger in any generation, the quantity $\phi_n(k, t)$, $k \in [n]$ is defined as follows: the probability that the ancestral sample is of size k at ancestral generation t with all mergers in prior backward WF steps being single binary mergers is $\phi_n(k, t)$. The allowed values for k are $k = 1, \ldots, N(t)$ for $t > 0$. When $k = 0$, however, $\phi_n(k, t)$ has a special interpretation: $\phi_n(0, t)$ is the probability that the WF genealogy from the current generation to ancestral generation t includes something other than a single binary merger in some generation. When $t = 0$, the algorithm is initialized using $\phi_n(n, 0) = 1$ and $\phi_n(k, 0) = 0$ for $k \neq n$, and in particular, $\phi_n(k, 0) = 0$ for $0 \leq k \leq n-1$.

Suppose the data at time t is $\phi_n(k, t)$ with $k \in [n] \cup \{0\}$. The crux of the algorithm is to generate data at time $t+1$, and the recurrence

$$\phi_n(k, t+1) = \sum_{\ell=k}^{\ell=k+1} \phi_n(\ell, t)p(\ell - k, \ell, N(t+1)) \tag{2}$$

does that for $k = 1, \ldots, \min(n, N(t+1))$. If the size of the ancestral sample in generation $t+1$ is k, the ancestral sample size in generation t must be either $\ell = k$ or $\ell = k+1$ because simultaneous binary mergers and triple mergers are precluded by the definition of $\phi_n(k, t+1)$. The two possibilities are disjoint, and the recurrence sums over the two possibilities. The recurrence for $\phi_n(k, t)$ is similar in structure to Eq. (3) in the appendix of Bhaskar et al. (2014).[2] The recurrences for genealogical quantities (as well as for the sample frequency spectrum viewed from a genealogical perspective) generally have a similar form (Tavaré, 1984).

The quantity $\phi_n(0, t+1)$, which has a special interpretation, is calculated using

$$\phi_n(0, t+1) = 1 - \phi_n(1, t+1) - \cdots - \phi_n(n^*, t+1),$$

where $n^* = \min(n, N(t+1))$.

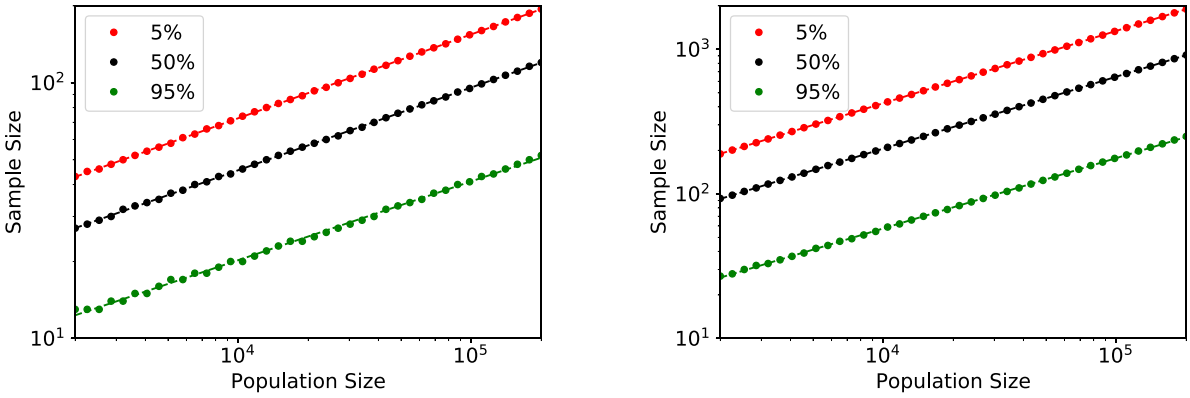
The algorithm is terminated at the tth ancestral generation if $\phi_n(0, t) + \phi_n(1, t) > 1 - 10^{-4}$. At termination, the probability that the sample has either coalesced to a single ancestral sample or some backward WF step involves a merger other than a single binary merger is greater than 0.9999.

The probability of a simultaneous binary merger or a triple merger between ancestral generations t and $t+1$ conditioned on at most a single binary merger in any backward WF step from 0 to t is

$$\frac{\phi_n(0, t+1) - \phi_n(0, t)}{1 - \phi_n(0, t)}. \tag{3}$$

This formula is used to visualize the effect of bottlenecks.

[2] An alternative version of Eq. (3) of Bhaskar et al. (2014) is as follows. In their notation, the recurrence can be written as $\mathbb{E}M^{(t)}_{n,k} = \sum_{m=1}^{n} p^{(t)}_{n,m}\left(\mathbb{E}M^{(t+1)}_{m,k} + mf(n, m, k)\right)$, where $f(n, m, k) = \frac{\binom{n}{k}\left\{{n-k \atop m-1}\right\}}{m\left\{{n \atop m}\right\}}$ is the probability that a given bin has exactly k balls when n balls are assigned to m bins randomly, conditioned on each bin receiving at least one ball.



(a) Only single binary mergers allowed.

(b) No triple mergers allowed.

Fig. 1. Probability of coalescence under WF with at most a single binary merger per generation and, alternatively, with no triple merger in any generation for various constant population sizes. In each plot, the sample sizes at which the probability is 5%, 50%, and 95% are shown as solid circles. The dashed lines are linear fits.

Probability of no triple merger

The algorithm to calculate the probability of no triple merger in the WF genealogy of a sample of size n is similar. The quantity $\psi_n(k,t)$, $k \in \mathbb{Z}^+$, is defined as follows: $\psi_n(k,t)$ is the probability that the ancestral sample is of size k in ancestral generation t with no triple mergers between generation 0 and ancestral generation t. As before, the definition of $\psi_n(0,t)$ is special: $\psi_n(0,t)$ is the probability of a triple merger in the WF genealogy between generation 0 and ancestral generation t. Again as before, the algorithm is initialized using $\psi_n(n,0) = 1$ and $\psi_n(k,0) = 0$ for $0 \le k \le n-1$.

Suppose the data at time t is $\psi_n(k,t)$ with $k \in [n] \cup \{0\}$. The recurrence

$$\psi_n(k,t+1) = \sum_{\ell=k}^{\ell=\min(n,N(t),2k)} \psi_n(\ell,t)p(\ell-k,\ell,N(t+1)) \qquad (4)$$

calculates data at $t+1$ for $k = 1,\ldots,\min(n,N(t+1))$. If the ancestral sample size at $t+1$ is k, the ancestral sample size at t, which is denoted by ℓ, must be at least k. It can be at most $2k$ because any backward WF step that whittles down a sample of size greater than $2k$ to k must involve a triple merger. In addition, ℓ cannot exceed n or $N(t)$. The recurrence is obtained by summing over all possibilities for ℓ. As before,

$$\psi_n(0,t+1) = 1 - \psi_n(1,t+1) - \cdots - \psi_n(n,t+1),$$

and we stop calculating when $\psi_n(0,t) + \psi_n(1,t) > 1 - 10^{-4}$.

The probability that there is a triple merger in the backward WF step from t and $t+1$ conditioned on no triple merger from 0 to t is

$$\frac{\psi_n(0,t+1) - \psi_n(0,t)}{1 - \psi_n(0,t)}. \qquad (5)$$

Like (3), this formula is also used to visualize the effect of bottlenecks.

This algorithm (and analogously the earlier algorithm) can be sped up by ignoring $\psi_n(k,t)$ if $\psi_n(k,t) < \epsilon_{tol}$ for an ϵ_{tol} that is small. As it is, the algorithm would maintain the probabilities $\psi_n(k,t)$ for $k \in [n] \cup \{0\}$ typically. As t increases, a probability such as $\psi_n(n,t)$ becomes quite small but can still remain positive. Holding on to such tiny numbers makes the algorithm quite expensive for large sample sizes. If probabilities smaller than ϵ_{tol} are ignored, there is a rapid reduction in the sample sizes that are tracked at ancestral generation t in the initial stages of the algorithm if n is large. The total contribution of $\psi_n(\ell,t)$ to probabilities in all later stages is bounded by $\psi_n(\ell,t)$ because the recurrence sums over disjoint possibilities. Suppose all probabilities smaller than ϵ_{tol} are ignored. The total probability ignored is then bounded by $\epsilon_{tol}nG$, where n is the sample size and G is the total number of generations. We use $\epsilon_{tol} = 10^{-120}$ so that the ignored probability is vanishingly small even with $n = G = 10^{20}$.

4. Verification and visualization

The algorithms for calculating the probabilities of at most a single binary merger in any generation of the WF genealogy and of no triple merger in any generation enable a direct verification of the asymptotic theory. Fig. 1 shows calculations for various population sizes. For each population size, the sample sizes at which the probabilities of coalescence with only single binary mergers (plot (a)) or with no triple mergers (plot (b)) are 5%, 50%, and 95% are shown.

Evidently, a higher sample size implies a higher probability of triple mergers and of more than a single binary merger in some generation in the WF genealogy of the sample. Sample sizes for which probabilities of coalescence with only single binary mergers are 5%, 50%, and 95% may be fitted as

$$3.55 \times N^{0.33},\ 2.31 \times N^{0.32},\ 1.19 \times N^{0.31},$$

respectively. The quality of the fit is quite good for N as small as 1000. The exponents are close to 1/3 as predicted by the asymptotic theory.

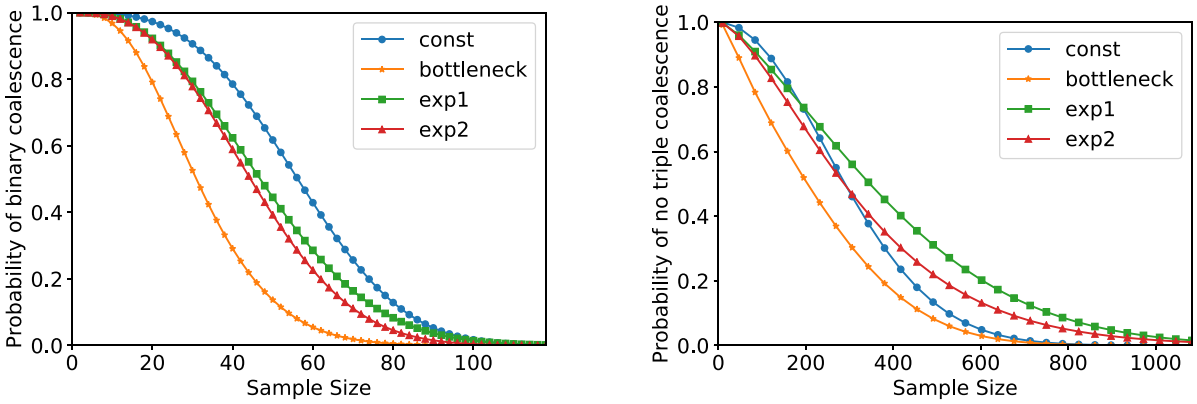
The linear fits for the no triple merger case are in even better agreement with the asymptotic theory. In this case, the sample sizes for which the probabilities of no triple merger are 5%, 50%, and 95% are

$$4.23 \times N^{0.50},\ 2.11 \times N^{0.50},\ 0.65 \times N^{0.49},$$

respectively. The exponents are close to 1/2 as predicted by the asymptotic theory. To increase the probability of WF coalescence with no triple merger from 5% to 95% the sample size needs to be decreased by a factor of six approximately.

Both algorithms allow for variable population sizes. The four demographic models of human population we consider are the same as in Bhaskar et al. (2014). These models are:

- Constant population with $N = 2 \times 10^4$, which is the baseline assumption in human genetics (Durrett, 2008).
- Constant population with $N(t) = 2 \times 10^4$ except for two bottlenecks: the first being $620 < t \le 720$ with $N(t) = 1000$ and the second being $4620 < t \le 4720$ with $N(t) = 300$, which is a drop-off by nearly a factor of 100. This model is based on Keinan et al. (2007).



(a) Only single binary mergers allowed.

(b) No triple mergers allowed.

Fig. 2. Probabilities of at most a single binary merger in any generation of the WF genealogy and, alternatively, of no triple merger in any generation for four demographic models and various sample sizes.

- Exponential decay for $0 \leq t \leq 920$ from $N(0) = 7 \times 10^4$ to $N(920) = 2 \times 10^3$, followed by $N(t) = 4000$ for $920 < t \leq 2000$, followed by $N(t) = 30{,}000$ for $2000 < t \leq 5900$, and $N(t) = 13{,}000$ for $t > 5900$. This model is based on Gravel et al. (2011). This model features a single exponential and is labeled `exp1` in Fig. 2.
- Exponential decay for $0 \leq t \leq 214$ from $N(0) = 10^6$ to $N(214) = 2 \times 10^4$, exponential decay for $214 \leq t \leq 920$ with $N(920) = 2050$, $N(t) = 4000$ for $920 < t \leq 2000$, $N(t) = 3 \times 10^4$ for $2000 < t \leq 5900$, and $N(t) = 13{,}000$ for $t > 5900$. This model features two exponentials and is therefore labeled `exp2` in Fig. 2. This model is based on Tennessen et al. (2012).

Fig. 2 shows that the probabilities of triple mergers and of something other than single binary mergers in WF genealogies increase noticeably because of bottlenecks.

Figs. 3 and 4 give a more explicit visualization of the effect of bottlenecks. In Fig. 3b, the distribution of possible ancestral sample sizes, conditioned on at most a single binary merger in prior generations, noticeably shifts downwards when the first bottleneck is encountered. The conditional probability of something other than a simultaneous binary merger or a triple merger in the backward WF step from t to $t+1$, as given by (3), spikes at the first bottleneck. At the second bottleneck, there is no such prominent spike. However, the distribution of possible ancestral sample sizes, allowing only single binary mergers, noticeably shifts downwards at the second bottleneck, even though the bottleneck is more than 4500 ancestral generations away and the sample size is only 100.

Our interpretation of the phenomena in Fig. 3(c) and (d) is as follows. In both cases, the heat-maps of $\phi_n(k, t)$ show evidence of an inflection point. In these models with exponential decay in ancestral population sizes, there is less pressure on the sample to shrink initially. However, the exponential decay appears to eliminate that effect at the inflection point. In both plots, the spike in the conditional probability given by (3) appears to be located near the inflection point.

In Fig. 4, the same phenomena are in evidence. Some of the phenomena are a little more prominent here. For example, a small spike in the conditional probability given by (5) is visible even at the second bottleneck in part (b) of the figure.

5. Discussion

The roots of the Kingman coalescent may be found in the work of Ewens (1972) and Watterson (1975). It was derived (Kingman, 1982a, b) at a time when a whole genome was yet to be sequenced and sample sizes did not go much beyond 10. Thus, it was natural to prove its convergence assuming the sample size to be fixed and small.

Data sets with more than 10^4 samples are now publicly available (Karczewski et al., 2016; Sudlow et al., 2015). Thus, it is essential to consider a convergence theory that does not fix the sample size, as we have done here.

The convergence theory we have developed is with reference to the Kingman partition distribution (see Theorem 1). If the current sample size is n and the ancestral sample is of size k, the ancestral sample induces a partition of the set $[n]$ into k subsets, the distribution of which is given by the Kingman partition distribution. The Kingman partition distribution, therefore, captures the structure of the genealogical tree in complete detail, except for inter-coalescence times which are determined independently.

Statistics that are used in analyzing sequence data are considerably less refined. For example, the sample frequency spectrum partitions the current sample into only two sets. We have proved that the WF sample frequency spectrum converges to that of the coalescent for samples of size $N^{1/3-\epsilon}$ or smaller. However, the $N^{1/3}$ bound on sample sizes is probably far from sharp because the proof proceeds via the Kingman partition distribution. A separate analysis of summary statistics such as the sample frequency spectrum would therefore be desirable.

Acknowledgments

The authors thank the reviewers as well as the editor for their valuable comments and Simon Tavaré for a useful pointer to the literature.

Appendix

This appendix gives proofs of theorems that were stated in the text. Statements of theorems are repeated in the interest of readability.

Theorem 1 (*Kingman, 1982b*)**.** *Suppose that the coalescent is run until the partition of* $[n]$ *consists of exactly* k *sets. If* $|A_j| = n_j$ *is the cardinality of* A_j*, the probability that the partition into* k *sets is* $\{A_1, \ldots, A_k\}$ *is equal to*

$$\frac{(n-k)!k!(k-1)!}{n!(n-1)!} n_1! n_2! \ldots n_k!.$$

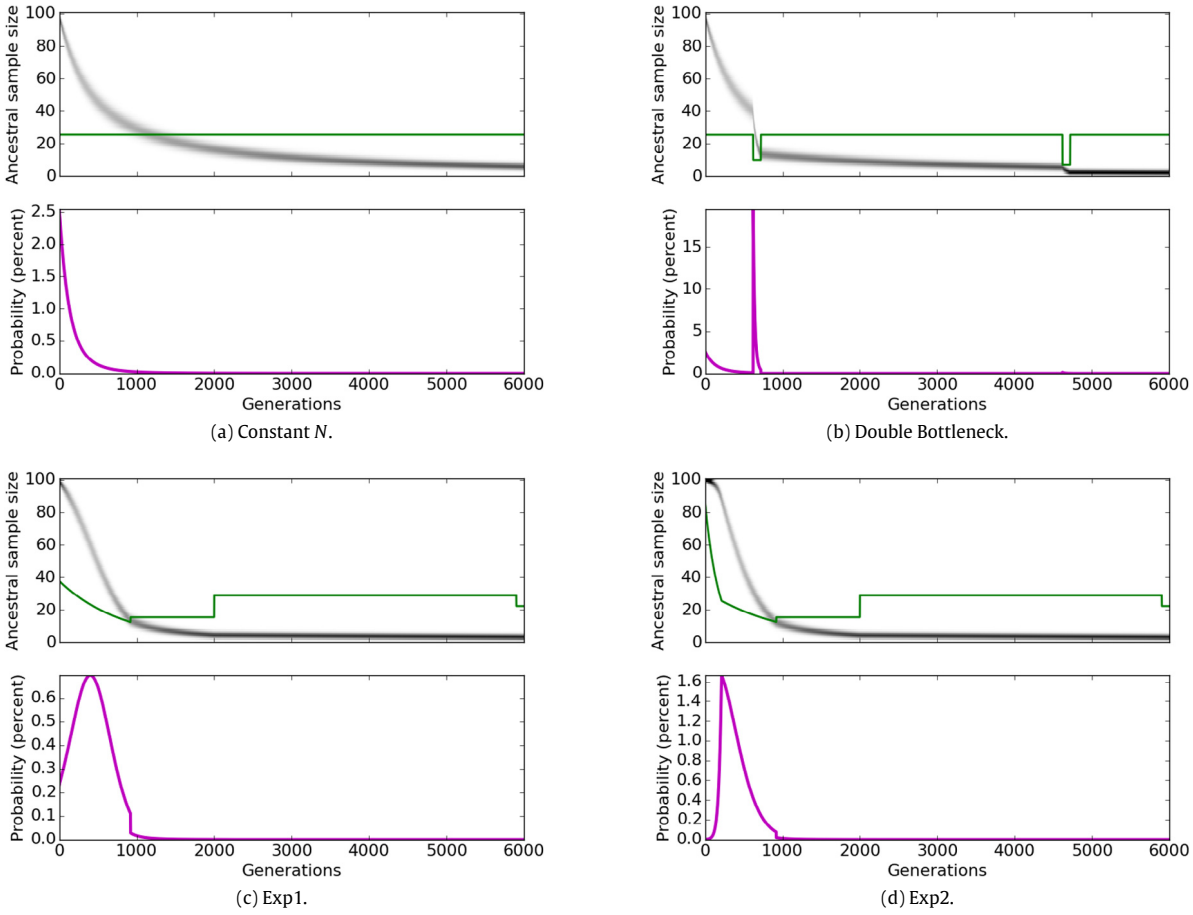


Fig. 3. The upper panels in (a) through (d) are heat-maps of probabilities $\phi_n(k, t)$, with black being 1 and white 0. The green line is a graph of $1.19 \times N(t)^{0.31}$. The lower panels in (a) through (d) graph the conditional probability given by (3). The plots (a) through (d) correspond to four different demographic models. The sample size is $n = 100$ in all the plots.

Proof. Because each coalescence is a union of two disjoint subsets of $[n]$, the coalescent process can be depicted as a forest of binary trees with each vertex a subset of $[n]$ and with the leaves being $\{1\}, \ldots \{n\}$. If disjoint subsets S_1 and S_2 coalesce, then $S_1 \cup S_2$ occurs as a vertex with S_1 and S_2 as its two children. Coalescences deeper into the ancestry are placed higher to capture the ordering of events. The leaves are lowest, and no two interior vertices occur at the same height. Because the Kingman coalescent is memoryless, every coalescent tree with the same root is generated with the same probability.

The number of coalescent trees with root A_1 and with their n_1 leaves being equal to $\{j\}$ for $j \in A_1$ is equal to $\frac{n_1!(n_1-1)!}{2^{n_1-1}}$. That is because the first union is any one of $n_1(n_1-1)/2$ possibilities, the second union any one of $(n_1-1)(n_1-2)/2$ possibilities, and so on. The total number of coalescence events in any of these trees is $n_1 - 1$. Likewise, the number of coalescent trees with root A_j is $\frac{n_j!(n_j-1)!}{2^{n_j-1}}$ and the number of coalescence events in any of these trees is $n_j - 1$.

The total number of forests with roots equal to $A_1, \ldots, A_k$ is equal to

$$\prod_{j=1}^{k} \frac{n_j!(n_j-1)!}{2^{n_j-1}}.$$

Although the order of coalescence events within a single tree is determined, the order of events between different trees is not determined. Because the number of coalescence events in a tree with root A_j is $n_j - 1$, the coalescence events corresponding to any given forest can be ordered in

$$\frac{\left(\sum_{j=1}^{k}(n_j-1)\right)!}{\prod_{j=1}^{k}(n_j-1)!} = \frac{(n-k)!}{\prod_{j=1}^{k}(n_j-1)!}$$

ways. Thus, the total number of sequences of $n-k$ coalescence events resulting in a forest with roots $A_1, \ldots, A_k$ is equal to $\frac{(n-k)!}{2^{n-k}}\prod_{j=1}^{k} n_j!$. Each sequence of $(n-k)$ coalescence events is equally likely by the memoryless property of the coalescent, and the total number of sequences of length $n-k$ is equal to $\frac{n!(n-1)!}{2^{n-k}k!(k-1)!}$, which implies the stated theorem. □

Corollary 2. *Suppose the set* $\{\{1\}, \ldots, \{n\}\}$ *undergoes* k *coalescences resulting in a partition of* $[n]$ *into* $n-k$ *sets. The probability* $q(k, n)$ *that each set in the resulting partition is of size* 1 *or* 2 *is given by* $q(k, n) = \frac{(n-k)^{\underline{k}}}{(n-1)^{\underline{k}}}$. *If* $3k \le n$ *and* $k \ge 2$, *we have* $\exp\left(-\frac{k^2}{2n}\right) \ge q(k, n) \ge \exp\left(-\frac{7k^2}{n}\right) \ge 1 - \frac{7k^2}{n}$.

Proof. The probability $q(k, n)$ is zero if $2k \ge n+1$ because a partition of size 3 or more is inevitable after so many coalescences. The formula for $q(k, n)$ is easily verified in this case.

Now suppose $2k \le n$. If a partition into $n-k$ sets has only sets of sizes 1 and 2, the number of sets of sizes 1 and 2 must be $(n-2k)$ and k, respectively. The number of such partitions is equal to

$$\binom{n}{2k}\frac{(2k)!}{2^k k!}.$$

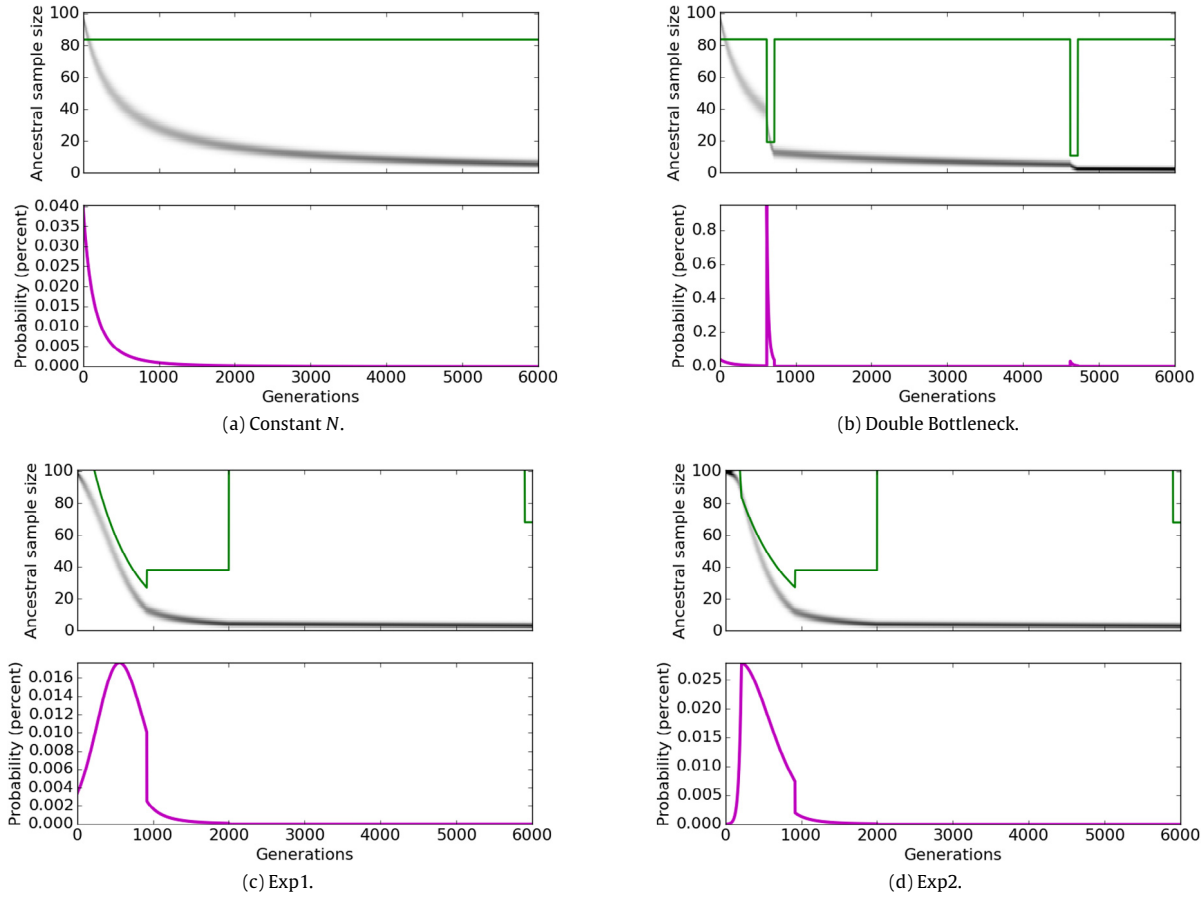


Fig. 4. The upper panels in (a) through (d) are heat-maps of probabilities $\psi_n(k, t)$, with black being 1 and white 0. The green line graphs $0.65 \times N(t)^{0.49}$. The lower panels in (a) through (d) graph the conditional probability given by (5). As before, (a) through (d) correspond to four different demographic models with sample size $n = 100$.

By Theorem 1, the probability of each partition is equal to

$$\frac{k!(n-k)!(n-k-1)!}{n!(n-1)!}2^k.$$

The proof of the formula for $q(k, n)$ is completed by multiplying the two numbers and simplifying.

The stated bounds for $q(k, n)$ follow from calculations that are elementary but a little tedious.

Let $p = q(k, n)$. To bound p, note that $\log(1-\alpha) = -\alpha + \alpha^2 \int_0^1 \frac{-(1-t)}{(1-\alpha t)^2}\, dt$ for $|\alpha| < 1$. If $\alpha \in [0, \frac{1}{2}]$, we may deduce that

$$\log(1-\alpha) = -\alpha - u\alpha^2 \tag{6}$$

for some $u \in [\frac{1}{2}, 1]$. By similar arguments based on the Euler summation formula,

$$\frac{1}{m} + \frac{1}{m+1} + \cdots + \frac{1}{n-1} = \log\left(\frac{n}{m}\right) + u\left(\frac{1}{m} - \frac{1}{n}\right) \tag{7}$$

and

$$\frac{1}{m^2} + \frac{1}{(m+1)^2} + \cdots + \frac{1}{(n-1)^2} = \left(\frac{1}{m} - \frac{1}{n}\right) + u\left(\frac{1}{m^2} - \frac{1}{n^2}\right), \tag{8}$$

for $m, n \in \mathbb{Z}^+$, $m < n$, and some $u \in [0, 1]$.

From the formula for $q(k, n)$ and (6), we have

$$\begin{aligned}\log p &= \sum_{j=1}^{k-1} \log\left(1 - \frac{k}{n-j}\right) \\ &= -\sum_{j=1}^{k-1} \frac{k}{(n-j)} - u\sum_{j=1}^{k-1} \frac{k^2}{(n-j)^2}\end{aligned}$$

for some $u \in [\frac{1}{2}, 1]$. The application of (6) is justified because $3k \le n$ implies $k/(n-k+1) < 1/2$. Applying (7) and (8), we get

$$\begin{aligned}\log p = &-k\log\left(\frac{n}{n-k+1}\right) - u_1 k\left(\frac{1}{n-k+1} - \frac{1}{n}\right) \\ &- u_2 k^2\left(\frac{1}{n-k+1} - \frac{1}{n}\right) \\ &- u_3 k^2\left(\frac{1}{(n-k+1)^2} - \frac{1}{n^2}\right)\end{aligned}$$

for some $u_1 \in [0, 1]$, $u_2 \in [\frac{1}{2}, 1]$, and $u_3 \in [0, 1]$.

Thus,

$$\begin{aligned}\log p &\ge k\log\left(1 - \frac{k-1}{n}\right) - \frac{k}{n-k+1} \\ &\quad - k^2\left(\frac{1}{n-k+1} + \frac{1}{(n-k+1)^2}\right) \\ &\ge -\frac{k(k-1)}{n} - \frac{k(k-1)^2}{n^2} - \frac{k}{n-k+1}\end{aligned}$$

$$
\begin{aligned}
&- k^2\left(\frac{1}{n-k+1}+\frac{1}{(n-k+1)^2}\right)\\
&\geq -\frac{k^2}{n}-\frac{k^3}{n^2}-\frac{3k}{2n}-k^2\left(\frac{3}{2n}+\frac{9}{4n^2}\right)\\
&\geq -\frac{7k^2}{n},
\end{aligned}
$$

where the second inequality is obtained using (6). We then have $p \geq \exp(-7k^2/n) \geq 1 - 7k^2/n$, proving the lower bound.

To prove the upper bound, argue as follows:

$$
\begin{aligned}
\log p &\leq k\log\left(1-\frac{k-1}{n}\right)-\frac{k^2}{2}\left(\frac{1}{n-k+1}-\frac{1}{n}\right)\\
&\leq -\frac{k(k-1)}{n}-\frac{k^2}{2(n-k+1)}+\frac{k^2}{2n}\\
&= -\frac{k^2}{2n}+\frac{k}{n}-\frac{k^2}{2(n-k+1)}\\
&\leq -\frac{k^2}{2n},
\end{aligned}
$$

where the second inequality uses $\log(1-x) \leq -x$ for $x \in (0, 1)$. The last inequality requires $k \geq 2$. □

Lemma 3. *Consider the application of a single backward WF step to a sample of size n with parental population of size N. Let p_d be the conditional probability that there is a single binary merger given that there is some merger. Then*

$$p_d \geq 1-\frac{(n-2)(n-1)}{2N}.$$

Proof. Let A_{12} be the event that samples 1 and 2 merge under the backward WF step, more specifically 1 and 2 have the same WF parent. Obviously, $\mathbb{P}(A_{12}) = \frac{1}{N}$.

Let $A_{12}^{(t)}$ be the event that 1 and 2 merge and that one of the other $(n-2)$ samples has the same parent as 1 and 2, implying a triple merger or worse. We have $A_{12}^{(t)} = \cup_{j=3}^{n} A_{12j}$, where A_{12j} is the event where 1, 2, and j have the same parent. Because $\mathbb{P}(A_{12j}) = \frac{1}{N^2}$, we have $\mathbb{P}(A_{12}^{(t)}) \leq \frac{(n-2)}{N^2}$.

Let $A_{12}^{(d)}$ be the event that 1 and 2 merge and that there is some other pair that merges, implying two binary mergers or worse. We have $A_{12}^{(d)} = \cup_{j,k} A_{12,jk}$, where $A_{12,jk}$ is the event that 1, 2 as well as j, k have the same parent. The union is over $2 < j < k \leq n$. Because $\mathbb{P}(A_{12,jk}) = \frac{1}{N} \times \frac{N-1}{N} \times \frac{1}{N} \leq \frac{1}{N^2}$, we have $\mathbb{P}(A_{12}^{(d)}) \leq \binom{n-2}{2}\frac{1}{N^2}$.

If $\tilde{A}_{12}$ is the event that 1, 2 merge and there is no other merger, $\tilde{A}_{12} = A_{12} - A_{12}^{(t)} - A_{12}^{(d)}$. Therefore,

$$
\begin{aligned}
\mathbb{P}(\tilde{A}_{12}) &\geq \frac{1}{N}-\frac{(n-2)}{N^2}-\binom{n-2}{2}\frac{1}{N^2}\\
&= \frac{1}{N}\left(1-\frac{(n-2)(n-1)}{2N}\right).
\end{aligned}
$$

The probability that there is a single binary merger during a backward Wright-Fisher step is equal to $\binom{n}{2}\mathbb{P}(\tilde{A}_{12})$.

If $\mathcal{C}$ is the event that there is some merger, $\mathcal{C} = \cup_{j,k} A_{jk}$, union over $1 \leq j < k \leq n$. Therefore, $\mathbb{P}(\mathcal{C}) \leq \binom{n}{2}\frac{1}{N}$. The lower bound for p_d in the lemma is obtained by simplifying

$$\frac{\binom{n}{2}\mathbb{P}(\tilde{A}_{12})}{\binom{n}{2}\frac{1}{N}}.$$ □

Theorem 4. *Each backward WF step in the genealogy of a sample of size $N^{1/3-\epsilon}$, $\epsilon > 0$, includes at most a single binary merger with probability converging to 1 as $N \to \infty$.*

Proof. Let $\mathcal{D}$ be the event that a sample of size n undergoes more than a single binary merger in some backward WF step in its genealogy. Let $\mathcal{D}_k$ be the event that the ancestral sample size is equal to k in some generation but the ancestral sample size is never $k-1$. Evidently, $\mathbb{P}(\mathcal{D}) \leq \sum_{k=3}^{n}\mathbb{P}(\mathcal{D}_k)$.

By Lemma 3,

$$
\begin{aligned}
\mathbb{P}(\mathcal{D}_k) &= \mathbb{P}\left(\text{ancestral sample never of size } k-1 \middle| \text{sample of size } k\right)\\
&\quad \times \mathbb{P}\left(\text{ancestral sample is of size } k \text{ in some generation}\right)\\
&\leq \frac{(k-2)(k-1)}{2N}.
\end{aligned}
$$

Therefore,

$$\mathbb{P}(\mathcal{D}) \leq \sum_{k=3}^{n}\frac{(k-3)(k-1)}{2N} \leq \frac{n^3}{2N}.$$

If $n = N^{1/3-\epsilon}$, $\mathbb{P}(\mathcal{D}) \leq N^{-3\epsilon}/2$, which converges to zero as $N \to \infty$. In the complement of $\mathcal{D}$, every merger in a backward WF step is a single binary merger. □

Let $\mathcal{D}_n$ be the event that there are more than two binary mergers or a triple merger or worse when a backward WF step is applied to a sample of size n. Then

$$\mathbb{P}\left(\mathcal{D}_n \middle| 1 \text{ and } 2 \text{ merge}\right) \leq \frac{n-2}{N}+\binom{n-2}{3}\frac{1}{N^2}+3\binom{n-2}{4}\frac{1}{N^2},$$

where the first term accounts for any of the samples 3 through n having the same parent as 1 and 2, the second term accounts for triple mergers with a parent other than that of 1 or 2, and the third term accounts for the possibility that there are two or more additional binary mergers. This bound simplifies to

$$\mathbb{P}\left(\mathcal{D}_n \middle| 1 \text{ and } 2 \text{ collide}\right) \leq \frac{n}{N}+\frac{n^4}{4N^2}.$$

This is an almost correct bound for the conditional probability of $\mathcal{D}_n$ given *any* merger, as we may expect from the high degree of symmetry. The argument below makes the idea rigorous by using more detailed conditioning.

The event $\mathcal{C}_{j,k}$, which we presently define and with respect to which we will condition later, pertains to a single backward WF step. The sample size is assumed to be n.

- Samples 1 through $j-1$ have unique parents and do not merge with any sample.
- The parent of sample j differs from the parents of samples $j+1, \ldots, k-1$.
- Samples j and k have the same parent.

Lemma 5. *Let p_1 be the conditional probability given $\mathcal{C}_{j,k}$ that some sample has the same parent as j and k. We have*

$$p_1 \leq \frac{n-k}{N-j+1} \leq \frac{n}{N-n}.$$

Proof. None of the samples $[k] - \{j, k\}$ are allowed to have the same parent as j and k subject to the condition $\mathcal{C}_{j,k}$. Subject to the condition $\mathcal{C}_{j,k}$, samples $k+1, \ldots, n$ can have any of $N-j+1$ parents (the $j-1$ parents of $1, \ldots, j-1$ are excluded). The probability of ending up with the same parent as j and k is thus $\frac{1}{N-j+1}$ for each of those $n-k$ samples, which proves the lemma. □

Lemma 6. *Let p_2 be the conditional probability given $\mathcal{C}_{j,k}$ that some three samples have the same parent and that parent is distinct from the parent of j and k. We have*

$$p_2 \leq \frac{n^3}{6(N-n)^2}.$$

Proof. The three samples of this lemma cannot belong to $[j] \cup \{k\}$. Thus, the three samples must be chosen out of a set of cardinality $n-(j+1)$, which can be done in

$$\binom{n-(j+1)}{3}$$

ways. For any such choice, the probability of a triple merger given $\mathcal{C}_{j,k}$ is

$$\frac{N-j}{N-j+1} \times \frac{1}{N-j+1} \times \frac{1}{N-j+1} \leq \frac{1}{(N-n)^2}.$$

The first factor accounts for the first member of the triple having to choose a parent other than those of $[j]$, and $1/(N-j+1)$ is the probability that the second or third member chooses the same parent as the first, subject to $\mathcal{C}_{j,k}$. Here, we have assumed that all three merging samples are chosen from $k+1, \ldots, n$. However, the same bound may be verified when one or more of the three is chosen from $j+1, \ldots, k-1$. Therefore,

$$p_2 \leq \binom{n-(j+1)}{3} \frac{1}{(N-n)^2} \leq \frac{n^3}{6(N-n)^2},$$

as claimed in the lemma. □

Lemma 7. *Let p_3 be the conditional probability given $\mathcal{C}_{j,k}$ that for each of c or more pairs, the two members of the pair have a common parent with that parent being distinct from the parents of all other pairs as well as the parent of j and k. We have*

$$p_3 \leq \frac{n^{2c}}{2^c c!(N-n)^c}.$$

Proof. The c pairs must be chosen out of the samples $[n]-[j]-\{k\}$. That means $n-(j+1)$ choices for each member of a pair and the samples which form c pairs can be chosen in

$$\binom{n-(j+1)}{2c}.$$

Having chosen the $2c$ samples, they can be paired in

$$(2c-1)(2c-3)\ldots 5.3.1 = \frac{(2c)!}{2^c c!}$$

ways because the first of the chosen samples can be paired in $2c-1$ ways following which the second of the remaining samples can be paired in $2c-3$ ways and so on. Having formed the pairs, the probability given $\mathcal{C}_{j,k}$ that each pair has a common parent distinct from that of other pairs as well as j and k is

$$\frac{(N-j)}{(N-j+1)^2} \times \frac{(N-j-1)}{(N-j+1)^2} \times \cdots \times \frac{(N-j-c+1)}{(N-j+1)^2} \leq \frac{1}{(N-n)^c}.$$

Here, we have assumed that all $2c$ samples that are paired are chosen from $k+1, \ldots, n$. The same bound may be verified when one or more of the samples is from $j+1, \ldots, k-1$. Therefore,

$$p_3 \leq \binom{n-(j+1)}{2c} \times \frac{(2c)!}{2^c c!} \times \frac{1}{(N-n)^c} \leq \frac{n^{2c}}{2^c c!(N-n)^c},$$

as claimed in the lemma. □

Lemma 8. *Let $\mathcal{D}_n$ denote the event that there are $c+1$ or more binary mergers with distinct parents or some triple merger in a single backward WF step applied to a sample of size n. Then*

$$\mathbb{P}\left(\mathcal{D}_n \middle| \mathcal{C}_{j,k}\right) \leq \frac{n}{N-n} + \frac{n^3}{6(N-n)^2} + \frac{n^{2c}}{2^c c!(N-n)^c}.$$

Proof. The event $\mathcal{D}_n \cap \mathcal{C}_{j,k}$ implies one of the following:

- Some sample has the same parent as j and k.
- Some three samples have a common parent distinct from the parent of j and k.
- There are c or more binary mergers in addition to the merger between j and k.

Therefore, $\mathbb{P}(\mathcal{D}_n|\mathcal{C}_{j,k}) \leq p_1 + p_2 + p_3$ proving the lemma. □

Theorem 9. *Each backward WF step in the genealogy of a sample of size $N^{\frac{c}{2c+1}-\epsilon}$, $\epsilon > 0$, includes at most c simultaneous binary mergers and no triple merger with probability converging to 1 in the limit of large N.*

Proof. Let $\mathcal{D}_\ell$ be the event that the ancestral sample size is ℓ and a backward WF step results in either a triple merger or more than c binary mergers. From the previous lemma,

$$\mathbb{P}(\mathcal{D}_\ell|\mathcal{C}_{j,k}) \leq \frac{\ell}{N-\ell} + \frac{\ell^3}{6(N-\ell)^2} + \frac{\ell^{2c}}{2^c c!(N-\ell)^c}.$$

Let $\mathcal{C}_\ell$ denote the event that a merger has occurred in a backward WF step with a sample size of ℓ. Evidently, $\mathcal{C}_\ell$ is the disjoint union of the events $\mathcal{C}_{j,k}$ over $1 \leq j < k \leq \ell$, with the event $\mathcal{C}_{j,k}$ asserting the first merger in lexicographic order is between sample j and k. Therefore,

$$\begin{aligned}\mathbb{P}\left(\mathcal{D}_\ell\middle|\mathcal{C}_\ell\right) &= \sum_{1\leq j<k\leq \ell} \mathbb{P}\left(\mathcal{D}_\ell\middle|\mathcal{C}_{j,k}\right)\mathbb{P}\left(\mathcal{C}_{j,k}\middle|\mathcal{C}_\ell\right) \\ &\leq \frac{\ell}{N-\ell} + \frac{\ell^3}{6(N-\ell)^2} + \frac{\ell^{2c}}{2^c c!(N-\ell)^c}.\end{aligned}$$

Let $\mathcal{D}$ be the event that a sample of size n undergoes either a triple merger or more than c binary mergers in some generation before coalescing to a single ancestor under WF. Then

$$\begin{aligned}\mathbb{P}(\mathcal{D}) &\leq \sum_{\ell=3}^{n} \mathbb{P}\left(\mathcal{D}_\ell\middle|\mathcal{C}_\ell\right) \\ &\leq \frac{(n+1)^2}{2(N-n)} + \frac{(n+1)^4}{24(N-n)^3} + \frac{n^{2c+1}}{2^c c!(N-n)^c}.\end{aligned}$$

The proof is completed by substituting $n = N^{\frac{c}{2c+1}-\epsilon}$ and verifying the $N \to \infty$ limit to be zero. □

We now turn to the sample frequency spectrum under WF. Unlike the approach in Griffiths and Tavaré (1998) and Bhaskar et al. (2014), our approach does not look at the internal structure of the genealogical tree.

Let $\mathcal{M}_n$ denote the condition that the genealogy of a sample of size n involves exactly one mutation under either Kingman or WF. Let $\mathcal{B}_n$ denote the condition that each backward WF step in the genealogy of a sample of size n involves at most single binary mergers.

Let $q(n, 2N)$ denote the probability of a single binary merger in a backward WF step applied to a sample of size n under the condition that there are no simultaneous binary mergers or triple mergers. Then

$$q(n,N) = \frac{1-\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{n-1}{N}\right)}{1-\mathfrak{c}_{n,N}},$$

where $\mathfrak{c}_{n,N}$ is the probability that a sample of size n either has a triple with a common parent (triple merger) or two pairs each with a common parent (simultaneous binary merger). Bounds for $q(n,N)$ will be given later. The probability of a mutation event in a single backward WF step is assumed to be $n\mu$. Given that either a mutation event or a coalescence event has occurred, the probability that it is a mutation is equal to

$$\frac{n\mu}{n\mu + q(n,N)}.$$

The probability it is a coalescence is equal to

$$\frac{q(n,2N)}{n\mu + q(n,N)}.$$

We are making the usual assumption that the sample cannot be hit with both a mutation and a merger in the same generation. The assumption could be unreasonable for large samples. However, we limit ourselves to samples of size $N^{1/3-\epsilon}$ or less. In addition, the condition $\mathcal{M}_n$ limits the total number of mutations in the genealogy of the sample to one, which makes the assumption reasonable even for large N.

The probability that a mutation strikes when the WF ancestral sample size is k but not when the sample size belongs to $[n]-\{1,k\}$ is equal to

$$\prod_{j=2}^{n} \frac{q(j,N)}{j\mu + q(j,N)} \times \frac{k\mu}{k\mu + q(k,N)}.$$

Therefore, conditioned on $\mathcal{M}_n \cap \mathcal{B}_n$, the probability that mutation strikes a sample of size n before any coalescence event is equal to

$$\frac{\frac{n\mu}{n\mu+q(n,N)}}{\sum_{j=2}^{n} \frac{j\mu}{j\mu+q(j,N)}}.$$

We take the limit $\mu \to 0$ to get

$$\mu_n = \frac{\frac{n}{q(n,N)}}{\sum_{j=2}^{n} \frac{j}{q(j,N)}}.$$

Thus, μ_n is the probability that a mutation is the first event to strike a sample of size n conditioned on $\mathcal{M}_n \cap \mathcal{B}_n$ in the limit of zero mutation.

Let $f(j,n)$ be the probability that j out of n samples are mutants under the condition $\mathcal{M}_n \cap \mathcal{B}_n$. The recurrence for $f(j,n)$ is

$$f(j,n) = \mu_n[j=1] + (1-\mu_n)\left(f(j,n-1)\left(1-\frac{j}{n-1}\right) + f(j-1,n-1)\frac{j-1}{n-1}\right). \tag{9}$$

In this recurrence, we have used Knuth's notation (Graham et al., 1994; Knuth, 1997) by which $[j=1]$ evaluates to 1 if $j=1$ and 0 otherwise.

To obtain the classical formula for the sample frequency spectrum, replace μ_n by

$$\tilde{\mu}_n = \frac{\frac{1}{n-1}}{\sum_{j=2}^{n} \frac{1}{j-1}},$$

which is obtained by taking $q(j,N) = j(j-1)/2N$ following the Kingman model and assuming $\mathcal{M}_n$. The exact solution of the recurrence

$$\tilde{f}(j,n) = \tilde{\mu}_n[j=1] + (1-\tilde{\mu}_n)\left(\tilde{f}(j,n-1)\left(1-\frac{j}{n-1}\right) + \tilde{f}(j-1,n-1)\frac{j-1}{n-1}\right) \tag{10}$$

is given by

$$\tilde{f}(j,n) = \frac{\frac{1}{j}}{\sum_{j=1}^{n-1} \frac{1}{j}}$$

for $j = 1,\ldots,n-1$.

Lemma 10. *For $n < \sqrt{N}$, $\frac{n(n-1)}{2N}\left(1-\frac{n^2}{2N}\right) \le q(n,N) \le \frac{n(n-1)}{2N}\left(1+\frac{n^4}{N^2}\right)$.*

Proof. The proof of Lemma 3 shows that $\mathbb{P}(\tilde{A}_{12}) \ge \frac{1}{N}\left(1-\frac{n^2}{2N}\right)$. The probability $\mathfrak{c}_{n,N}$ of a simultaneous binary merger or a triple merger is bounded by

$$\begin{aligned}\mathfrak{c}_{n,N} &\le \binom{n}{3}\frac{1}{N^2} + 3\binom{n}{4}\frac{1}{N^2} \\ &\le \frac{n^4}{3N^2}.\end{aligned}$$

The lower bound follows from $q(n,N) = \frac{n(n-1)}{2}\mathbb{P}\left(\tilde{A}_{12}\right)/(1-\mathfrak{c}_{n,N}) \ge \frac{n(n-1)}{2}\mathbb{P}\left(\tilde{A}_{12}\right)$. The upper bound follows from

$$\begin{aligned}q(n,N) &= \frac{n(n-1)}{2}\mathbb{P}\left(\tilde{A}_{12}\right) \\ &\le \frac{n(n-1)/(2N)}{1-\mathfrak{c}_{n,N}} \\ &\le \frac{n(n-1)/(2N)}{1-n^4/(3N^2)} \\ &\le \frac{n(n-1)}{2n}\left(1+\frac{n^4}{N^2}\right).\end{aligned}$$

The last inequality requires $n^4 < 2N^2$ which follows from $n < \sqrt{N}$. □

Lemma 11. *For $n < \sqrt{N}$, we have*

$$\tilde{\mu}_n\left(1-\frac{n^2}{2N}\right)\left(1-\frac{n^4}{N}\right) \le \mu_n \le \tilde{\mu}_n\left(1+\frac{n^2}{N}\right)\left(1+\frac{n^4}{N^2}\right).$$

Proof. If we use the definition of μ_n and write

$$\mu_n = \frac{\frac{n}{q(n,2N)}}{\sum_{j=2}^{n}\frac{j}{q(j,2N)}} = \frac{\frac{1}{n-1}(1-s_n)}{\sum_{j=2}^{n}\frac{1}{j-1}(1-s_j)}$$

after taking $q(j,N) = \frac{j(j-1)}{2N}(1-s_j)$, then by the previous lemma $s_j \in [-j^4/N^2, j^2/2N]$.

To obtain the lower bound in the lemma, use $s_j \ge -\frac{j^4}{N^2} \ge -n^4/N^2$ in the denominator, use $s_n \le n^2/2N$ in the numerator, and simplify using $\left(1+n^4/N^2\right)^{-1} > 1-\frac{n^4}{N^2}$.

To obtain the upper bound in the lemma, use $s_n \ge -\frac{n^4}{N^2}$ in the numerator, use $\frac{n^2}{2N} \ge \frac{j^2}{2N} \ge s_j$ in the denominator, and simplify using $\left(1-\frac{n^2}{2N}\right)^{-1} < \left(1+\frac{n^2}{N}\right)$ for $n^2 < N$. □

Lemma 12. *If $n \le N^{1/3-\epsilon}$, then*

$$\lim_{N\to\infty}\frac{1}{2}\sum_{j=1}^{n-1}\left|f(j,n)-\tilde{f}(j,n)\right| = 0.$$

Proof. Note that $|ab-\tilde{a}\tilde{b}| \le \left|a-\tilde{a}\right||b| + |b-\tilde{b}|\left|\tilde{a}\right|$. Subtracting (9) and (10), we get

$$\begin{aligned}\left|f(j,n)-\tilde{f}(j,n)\right| &\le |\tilde{\mu}_n-\mu_n|\left(f(j,n-1)\left(1-\frac{j}{n-1}\right) + f(j-1,n-1)\left(\frac{j-1}{n-1}\right)\right) \\ &+ \tilde{\mu}_n\left|f(j,n-1)-\tilde{f}(j,n-1)\right|\left(1-\frac{j}{n-1}\right) \\ &+ \tilde{\mu}_n\left|f(j-1,n-1)-\tilde{f}(j-1,n-1)\right|\left(\frac{j-1}{n-1}\right)\end{aligned}$$

for $j = 2,\ldots,n-1$. For $j=1$, there is an additional $|\mu_n-\tilde{\mu}_n|$ term.

Summing these inequalities, we have

$$\sum_{j=1}^{n-1}\left|f(j,n)-\tilde{f}(j,n)\right| \leq 3\left|\mu_n-\tilde{\mu_n}\right|+\tilde{\mu}_n\sum_{j=1}^{n-1}\left|f(j,n-1)-\tilde{f}(j,n-1)\right|.$$

The factor 3 in the above inequality follows because

$$\begin{aligned}&\sum_{j=1}^{n-1}\left(f(j,n-1)\left(1-\frac{j}{n-1}\right)+f(j-1,n-1)\left(\frac{j-1}{n-1}\right)\right)\\ &\leq \sum_{j=1}^{n-1} f(j,n-1)+f(j-1,n-1)\\ &\leq 2.\end{aligned}$$

The 2 is changed to 3 to allow for an additional $|\mu_n-\tilde{\mu}_n|$ in the $j=1$ case. Because $\tilde{\mu}_n<1$ for $n>2$ and $f(j,n)\equiv\tilde{f}(j,n)$ for $n=2$, we have

$$\sum_{j=1}^{n-1}\left|f(j,n)-\tilde{f}(j,n)\right| \leq 3\sum_{k=3}^{n}\left|\mu_k-\tilde{\mu}_k\right|.$$

The proof is now easily completed by an application of the previous lemma. □

Theorem 13. *Let $f_{WF}(k,n)$ be the probability that k out of n samples are mutants conditional on exactly one mutation in the WF genealogy of the sample. Then the total variation distance*

$$\frac{1}{2}\sum_{k=1}^{n-1}\left|f_{WF}(k,n)-\frac{1/k}{\mathcal{H}_{n-1}}\right|\to 0$$

for $n\leq N^{1/3-\epsilon}$, $\epsilon>0$, in the limit of zero mutation and large N.

Proof. By Theorem 4, the probability that any backward WF step produces a simultaneous binary merger or a triple merger converges to zero as $N\to\infty$. Thus, we may assume the condition $\mathcal{B}_n$ in the limit of large N and invoke the previous lemma and infer this theorem. □

References

Aldous, D., 1989. Probability Approximations via the Poisson Clumping Heuristic. Springer, New York.

Bhaskar, A., Clark, A.G., Song, Y.S., 2014. Distortion of genealogical properties when the sample is very large. Proc. Natl. Acad. Sci. 111, 2385–2390.

Chen, H., Chen, K., 2013. Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. Genetics 194, 721–736.

Chen, H., Hey, J., Chen, K., 2015. Inferring very recent population growth rate from population-scale sequencing data: using a large-sample coalescent estimator. Mol. Biol. Evol. 32, 2996–3011.

Davies, J.L., Simančík, F., Lyngsø, R., Mailund, T., Hein, J., 2007. On recombination-induced multiple and simultaneous coalescent events. Genetics 177, 2151–2160.

Durrett, R., 2008. Probability Models for DNA Sequence Evolution. Springer Science & Business Media.

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

Fu, Y., 2006. Exact coalescent for the Wright-Fisher model. Theor. Popul. Biol. 69, 385–394.

Graham, R.L., Knuth, D.E., Patashnik, O., 1994. Concrete Mathematics, second ed. Addison-Wesley, NJ.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D., Altshuler, D.L., et al., 2011. Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. 108, 11983–11988.

Griffiths, R.C., 2006. Coalescent lineage distributions. Adv. Appl. Probab. 38, 405–429.

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

Griffiths, R.C., Tavaré, S., 1998. The age of a mutation in a general coalescent tree. Stoch. Models 14, 273–295.

Karczewski, K.J., Weisburd, B., Thomas, B., et al., 2016. The ExAC browser: displaying reference data information from over 60000 exomes. Nucleic Acids Res. 45, D840–D845.

Keinan, A., Mullikin, J.C., Patterson, N., Reich, D., 2007. Measurement of the human allele frequency spectrum demostrates greater genetic drift in East Asians than in Europeans. Nature Genet. 39, 1251.

Kingman, J.F.C., 2000. Origins of the coalescent: 1974-1982. Genetics 156, 1461–1463.

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

Kingman, J.F.C., 1982b. The coalescent. Stochastic Process. Appl. 13, 235–248.

Knuth, D.E., 1997. The Art of Computer Programming, vol. 1, third ed. Addison-Wesley, NJ.

Möhle, M., 2000. Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models. Adv. Appl. Probab. 32, 983–993.

Möhle, M., Sagitov, S., 2001. A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29, 1547–1562.

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

Polanski, A., Szczesna, A., Garbulowski, M., Kimmel, M., 2017. Coalescence computations for large samples drawn from populations of time-varying sizes. PLoS One 12.

Sudlow, C., Gallacher, J., Allen, N., et al., 2015. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12 (e1001779).

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

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al., 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64–69.

Wakeley, J., King, L., Low, B.S., Ramachandran, S., 2012. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190, 1433–1445.

Wakeley, J., Takahashi, T., 2003. Gene genealogies when the sample size exceeds the effective size of the population. Mol. Biol. Evol. 20, 208–213.

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