The Validity of the Coalescent Approximation for Large Samples

Andrew Melfi and Divakar Viswanath

July 31, 2017

Department of Mathematics, University of Michigan (melfi/divakar@umich.edu).

Abstract

The Kingman coalescent, widely used in genetics, is known to be a good approximation when the sample size is small relative to the population size. In this article, we investigate how large the sample size can get without violating the coalescent approximation. If the haploid population size is $2N$, we prove that for samples of size $N^{1/3-\epsilon}$, $\epsilon > 0$, coalescence under the Wright-Fisher (WF) model converges in probability to the Kingman coalescent in the limit of large N. For samples of size $N^{2/5-\epsilon}$ or smaller, the WF coalescent converges to a mixture of the Kingman coalescent and what we call the mod-2 coalescent. For samples of size $N^{1/2}$ or larger, triple collisions in the WF genealogy of the sample become important. The sample size for which the probability of conformance with the Kingman coalescent is 95% is found to be $1.47 \times N^{0.31}$ for $N \in [10^3, 10^5]$, showing the pertinence of the asymptotic theory. The probability of no triple collisions is found to be 95% for sample sizes equal to $0.92 \times N^{0.49}$, which too is in accord with the asymptotic theory.

Varying population sizes are handled using algorithms that calculate the probability of WF coalescence agreeing with the Kingman model or taking place without triple collisions. For a sample of size 100, the probabilities of coalescence according to the Kingman model are 2%, 0%, 1%, and 0% in four models of human population with constant N, constant N except for two bottlenecks, recent exponential growth, and increasing recent exponential growth, respectively. For the same four demographic models and the same sample size, the probabilities of coalescence with no triple collision are 92%, 73%, 88%, and 87%, respectively. Visualizations of the algorithm show that even distant bottlenecks can impede agreement between the coalescent and the WF model.

Finally, we prove that the WF sample frequency spectrum for samples of size $N^{1/3-\epsilon}$ or smaller converges to the classical answer for the coalescent.

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]

[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).

Kingman's motivation in deriving the coalescent (Kingman, 1982a,b) 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 because at every coalescence exactly two ancestral samples are picked at random and deemed to have a common parent. That memoryless property is the chief reason for its simplicity and usefulness.

However, the coalescent eliminates some important biological effects when the sample size is large relative to the population size. For example, if 10 haploid individuals are known to have one of two parents, the split of the two individuals between the parents is binomial in the Wright-Fisher (WF) model as we would expect. However, the partition of 10 individuals into two induced by the coalescent is uniform (Durrett, 2008) and not binomial.

The WF model, with its non-overlapping generations, is of course imperfect. For a phenomenon as complicated as the propagation of genetic material between generations and in diverse species, every model will be unsatisfactory in some way. The ability to make useful inferences by adjusting a few parameters is of greater consequence than microscopic faithfulness to actual phenomena, which are anyway unknowable to some extent. In this regard the WF model and even more so the Kingman model, with population sizes as well as mutation and recombination rates as their key parameters, have been quite useful.

The WF model is more reliable than the Kingman model when the sample sizes are large relative to population size. As implied by the classic birthday problem and its variants (Aldous, 1989), two individuals in a sample of size $N^{1/2}$, assuming a fixed population size of $2N$, may be expected to have a common parent. 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. The situation where every coalescence is a single double collision is the key assumption of Kingman's coalescent. However, when the sample size is $N^{2/3}$, three samples may be expected to have a common parent implying a triple collision. For sample sizes in-between $N^{1/2}$ and $N^{2/3}$, there will be simultaneous double collisions in a single generation. Kingman's coalescent captures such simultaneous and multiple collisions as a succession of double collisions, each with no memory whatsoever of the previous collisions. Such an approximation introduces artifacts such as the uniform split of individuals between parents as we have already noted.

This paper derives theorems and algorithms that help delineate the regions of validity of Kingman's coalescent relative to WF. Thus, the genealogical coalescence process is considered here using Kingman's model as well as the WF model. The coalescent by itself will always refer to the Kingman model, although it will be explicitly referred to as the Kingman coalescent if there is room for ambiguity. Coalescence under the WF model is referred to as WF coalescence.

Existing literature on large sample effects may be divided into two categories. In the first category, the focus is on rates of coalescence and the number of ancestral samples as a function of the ancestral generation, and the Kingman model is assumed. Tavaré (1984) obtained formulas for the size of the ancestral sample (number of lineages) as a function of the ancestral generation, assuming fixed populations 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 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.

In the other category, the limitations of the coalescent are addressed explicitly using the WF model. Wakeley and Takahashi (2003) observed (relying on the earlier work of Fisher) that if the sample size is $2Nx$ with $x \in (0, 1)$ (the same authors tweaked the WF model to allow even $x \in (0, 2]$), the expected parental sample size is $2N(1 - e^{-x})$. From that observation, they derived approximate estimates of the size of the coalescent tree as well as the length of the extremal branches. They concluded that large samples lead to more singletons (by about 10%) in the sample frequency spectrum. Fu (2006) used Stirling numbers to derive an exact coalescent for WF and came to a similar conclusion. More recently, Bhaskar et al. (2014) derived recurrences to compute the sample frequency spectrum as well as expectations of ancestral sample sizes exactly under WF. Connections to that work will be indicated later.

The contribution of this paper includes theorems that give sample sizes for which the coalescent agrees with WF (asymptotically). Simultaneous double collisions may be less disruptive (Bhaskar et al., 2014, Davies et al., 2007) than triple and other multiple collisions, and we prove theorems that indicate sample sizes for which all WF coalescences would be simultaneous double collisions and no worse. Algorithms that handle varying population sizes are derived and applied to demographic models of human population. These algorithms are used to visualize how bottlenecks in the distant past can disrupt the validity of the coalescent approximation. We prove a theorem about convergence of the sample frequency spectrum under WF to the classical answer derived using the coalescent. The Python/C program that implements algorithms we derive is posted at `github.com/melfiand/lsample`.

Validity of the coalescent approximation for large samples

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 $[n]$ is the first part of the coalescent and 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 $2N$ WF generations.

The jump chain and the death process are independent, and in many of our arguments, the death process does not play any role at all. The death process governs the rates of coalescence, which can be adjusted independently, for example to allow for varying population size, and at any rate the correspondence of the rates to the WF model is only approximate.

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

Suppose that the coalescent is run until the partition of $[n]$ *consists of exactly* k *sets. If the* $|A_j| = n_j$ *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 italicized paragraphs are proved in the appendix. For the above statement, the appendix gives a combinatorial proof in the spirit of Griffiths and Lessard (2005). Kingman's proof is recursive (Kingman, 1982b, Durrett, 2008).

The 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). When considering the genealogical process, the k samples in a generation choose parents from their parental generation independently, with each member of the parental generation being equally likely to be chosen. The members of the parental generation which 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 the sequence of backward WF steps until an ancestral generation with a single ancestral sample is reached.

The coalescent approximation may be violated because of simultaneous double collisions or triple collisions. Any multiple collisions higher than triple, such as quadruple collisions, always include triple collisions. Simultaneous double collisions in backward WF steps may be less disruptive because they can be produced by the coalescent with appreciable probability, as shown by the following corollary:

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 double collisions are produced with an appreciable probability 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 $2N$ that allow only single double collisions in the WF genealogy of the sample (with high probability) as well as bounds that allow simultaneous double collisions.

For a constant population size of $2N$, the following theorem indicates sample sizes that ensure that in every backward WF step at most two individuals have a common parent:

The WF coalescent of a sample of size $N^{1/3-\epsilon}$*,* $\epsilon > 0$*, converges in probability to the coalescent in the limit of large* N*.*

This theorem does not consider rates of coalescence (encoded in the death process). The theorem only claims that the probability that there are either simultaneous double collisions or triple collisions 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 WF coalescence 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 below later.

Suppose we look for a bound on the sample size that ensures that every WF coalescence consists of either one or two double collisions. We then have the following theorem:

The WF genealogy of a sample of size $N^{2/5-\epsilon}$, $\epsilon > 0$, includes only one or two simultaneous double collisions in any ancestral generation with probability tending 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 double collisions rather than just 2. We have the following theorem:

The probability that the WF genealogy of a sample of size $N^{\frac{c}{2c+1}-\epsilon}$, $\epsilon > 0$, consists only of double collisions, with the number of double collisions 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 collisions kick in for sample sizes of the order $N^{1/2}$ or higher. If N is large and the sample is size is smaller than $N^{1/2-\epsilon}$, we may assume that all collisions in backward WF steps are simultaneous double collisions.

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 $f(k,n) = \frac{1/k}{\mathcal{H}_{n-1}}$ for $k = 1, \ldots, n-1$ in the limit of zero mutation rate. The following theorem shows that WF sample frequency spectrum converges to the classical answer for sample sizes smaller than $N^{1/3-\epsilon}$.

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

$$\frac{1}{2}\sum_{k=1}^{n-1} |f_{WF}(k,n) - f(k,n)| \to 0$$

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

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 double collisions or triple collisions is strictly greater than zero. Indeed, the probability of such events in the transition to the parental generation by a single backward WF step is strictly greater than zero. However, by a theorem stated above and proved in the appendix, the probability that the WF genealogy includes only single double collisions converges to 1 in the limit $N \to \infty$ if $n \leq N^{1/3-\epsilon}$, where N is the constant population size.

The theorem may be amended to apply to populations sizes $CN^{1/3-\epsilon}$ for $\epsilon > 0$ and some positive constant C. However, in the absence of a numerical value for the constant C, there would be no additional information.

A numerical constant can be produced by an algorithm that calculates the probability of the WF genealogy including only single double collisions. In this section, we derive such an algorithm. We derive another algorithm that calculates the probability that the genealogy of a sample of size n does not include any triple collisions. 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 collision in a single Wright-Fisher step. Then

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

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

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

for $0 \leq 2k \leq n$. The formula is valid for $n > 2N$. In fact, it is valid for any k with usual conventions about binomial coefficients (Graham et al., 1994) and with the assumption $n \in \mathbb{Z}^+$. The formula may be justified as follows. First, we may choose $2k$ samples to participate in k simultaneous double collisions 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 $\frac{1}{2N}$. 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 all collisions being single double collisions

For the current generation from which a sample of n is taken, we assume $t = 0$. Let $2N(t)$ be the haploid population size t ancestral generations ago. To calculate the probability that the WF genealogy of the sample has only single double collisions, the quantity $\pi(k, t)$, $k \in \{1, \ldots, n\}$ is endowed with the following interpretation: at ancestral generation t, the probability that the ancestral sample is of size k with all prior coalescences being single double collisions is $\pi(k, t)$. The allowed values for k are $k = 1, \ldots, 2N(t)$ for $t > 0$. When $k = 0$, however, $\pi(k, t)$ has a different interpretation: $\pi(0, t)$ is the probability that WF coalescence has produced something other than a single double collision prior to ancestral generation t. For $t = 0$, k can be anything, although the algorithm is initialized using $\pi(n, 0) = 1$ and $\pi(k, 0) = 0$ for $k \neq n$, and in particular, $\pi(k, 0) = 0$ for $0 \leq k \leq n-1$.

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

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

does that for $k = 1, \ldots, \min(n, 2N(t+1))$. 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 double collisions and triple collisions are precluded by the definition of $\pi(k, t+1)$. The two possibilities are disjoint, and the recurrence sums over the two possibilities.

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

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

where $n^* = \min(n, 2N(t+1))$.

The algorithm is terminated at the tth ancestral generation if $\pi(0,t)+\pi(1,t) > 1-10^{-4}$. At termination, the probability that the sample has either coalesced to a single ancestral sample or the WF genealogy has violated the requirement of only single double collisions is greater than 0.9999.

Probability of no triple collisions

The algorithm to calculate the probability of no triple collisions is similar. The quantity $\pi(k,t)$, $k \in \mathbb{Z}^+$, now has the following interpretation: $\pi(k,t)$ is the probability that the ancestral sample is of size k in ancestral generation t with no triple collision in any of the t backward WF steps from generation 0 to ancestral generation t. As before, the interpretation of $\pi(0,t)$ is different: $\pi(0,t)$ is the probability of a triple collision prior to ancestral generation t. Again as before, the algorithm is initialized using $\pi(n,0) = 1$ and $\pi(k,0) = 0$ for $0 \le k \le n-1$.

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

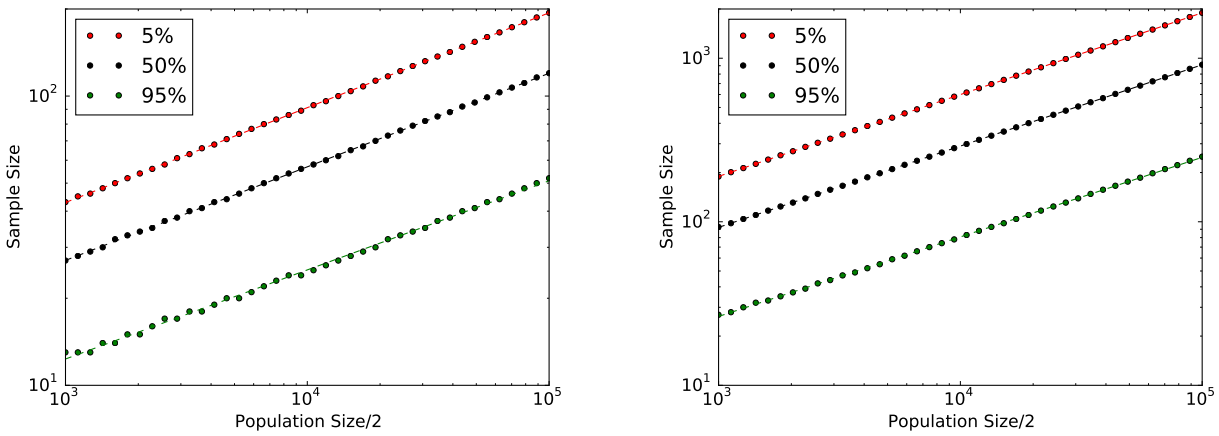
$$\pi(k, t+1) = \sum_{\ell=k}^{\ell=\min(n,2N(t),2k)} \pi(\ell, t) p(\ell - k, \ell, N(t+1))$$

calculates data at $t+1$ for $k = 1, \ldots, \min(n, 2N(t+1))$. 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 collision. In addition, ℓ cannot exceed n or $2N(t)$. The recurrence is obtained by summing over all possibilities for ℓ. As before,

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

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

This algorithm can be sped up by ignoring $\pi(k,t)$ if $\pi(k,t) < \epsilon_{tol}$ for an ϵ_{tol} that is small. As it is, the algorithm would maintain the probabilities $\pi(k,t)$ for $k \in [n] \cup \{0\}$ typically. As t increases, a probability such as $\pi(n,t)$ 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 $\pi(\ell,t)$ to probabilities in all later stages is bounded by $\pi(\ell,t)$ 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 $n = G = 10^{20}$. A similar device was employed by Bhaskar et al. (2014).



(a) Only single double collisions allowed.

(b) No triple collisions allowed.

Figure 1: Exact probability of coalescence with only single double collisions and no triple collisions, respectively, 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.

Verification and visualization

The algorithms for calculating the probability of WF coalescence with only single double collisions and no triple collisions enable a direct verification of the asymptotic theory. Figure 1 shows calculations for various population sizes. For each population size, the sample size at which the probability of coalescence according to the Kingman model (plot (a)) or with no triple collisions (plot (b)) are 5%, 50%, and 95% are also shown.

Evidently, a higher sample size implies a higher probability of triple collisions or of violating the Kingman model. Sample sizes for which probabilities of conformance with the Kingman model are 5%, 50%, and 95% may be fitted as

$$4.45 \times N^{0.33},\ 2.89 \times N^{0.32},\ 1.47 \times N^{0.31},$$

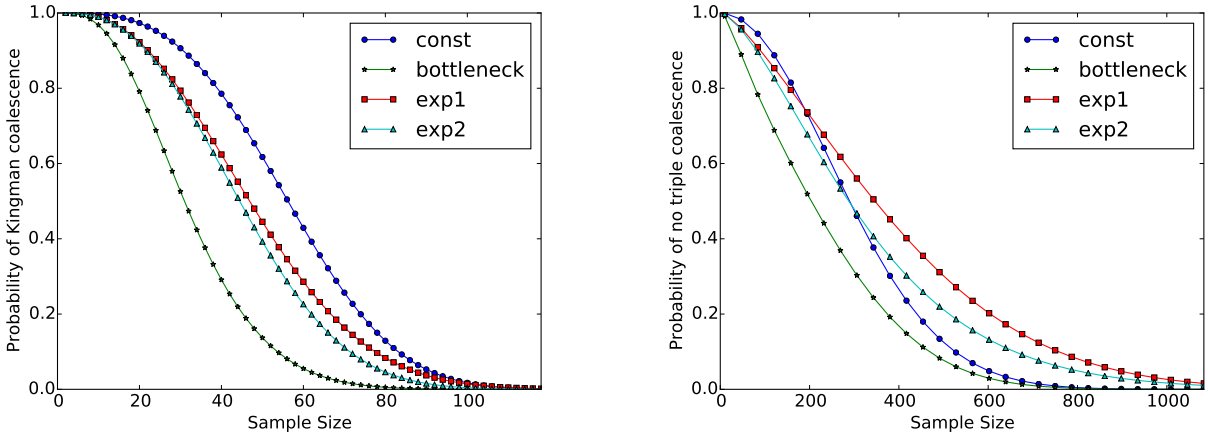
respectively. The quality of the fit is quite good for even N as small as 1000. The exponents are close to 1/3 as predicted by the asymptotic theory.

The quality fits for no triple collisions is even better. In this case, the sample sizes for which the probabilities of no triple collision are 5%, 50%, and 95% are

$$5.99 \times N^{0.50},\ 2.98 \times N^{0.50},\ 0.92 \times N^{0.49},$$

respectively. The exponents are close to 1/2 as predicted by the asymptotic theory. To increase the probability of WF coalescence with no triple collisions 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 apply the algorithms to are the same as in Bhaskar et al. (2014). These models are:



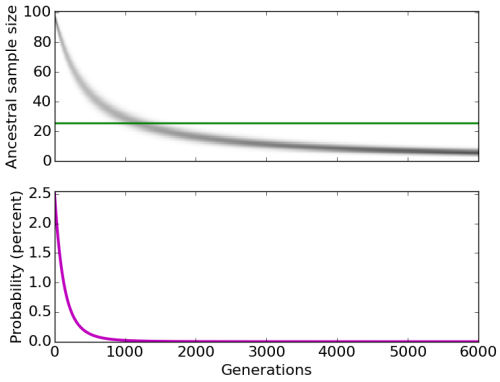
(a) (b)

Figure 2: Probabilities of coalescence with at most a single double collisions in any backward WF step and with no triple collisions for four demographic models and various sample sizes.

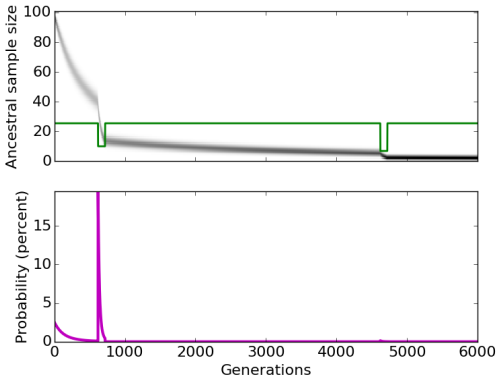
- Constant population with $N = 10^4$, which is the baseline assumption in human genetics (Durrett, 2008).
- Constant population with $N(t) = 10^4$ except for two bottlenecks: the first being $620 < t \leq 720$ with $N(t) = 500$ and the second being $4620 < t \leq 4720$ with $N(t) = 150$, which is a drop-off by nearly a factor of 100. This model is based on Keinan et al. (2007).
- Exponential decay for $0 \leq t \leq 920$ from $N(0) = 3.5 \times 10^4$ to $N(920) = 10^3$, followed by $N(t) = 2000$ for $920 < t \leq 2000$, followed by $N(t) = 15{,}000$ for $2000 < t \leq 5900$, and $N(t) = 6{,}500$ for $t > 5900$. This model is based on Gravel et al. (2011). This model features a single exponential and is labeled `exp1` in Figure 2.
- Exponential decay for $0 \leq t \leq 214$ from $N(0) = 5 \times 10^5$ to $N(214) = 10^4$, exponential decay for $214 \leq t \leq 920$ with $N(920) = 1025$, $N(t) = 2000$ for $920 < t \leq 2000$, $N(t) = 15000$ for $2000 < t \leq 5900$, and $N(t) = 6500$ for $t > 5900$. This model features two exponentials and is therefore labeled `exp2` in Figure 2.

Figure 2 shows that the probabilities of triple collision and of violation of the Kingman model increase noticeably because of bottlenecks.

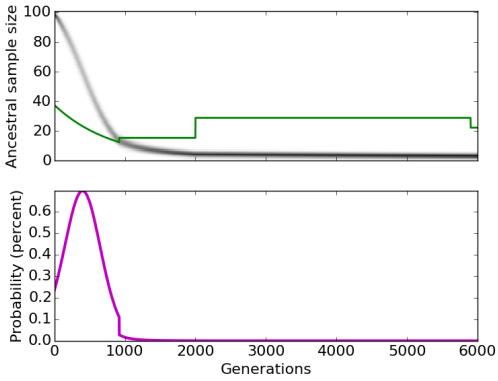
Figures 3 and 4 give a more explicit visualization of the effect of bottlenecks. In Figure 3b, the distribution of possible ancestral sample sizes, conditioned on no violation, noticeably shifts downwards when the first bottleneck is encountered. The conditional probability of a violation spikes at the first bottleneck. At the second bottleneck, there is no such prominent spike in the conditional probability of a violation. However, the distribution of possible ancestral sample sizes, conditioned on no violation, 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.



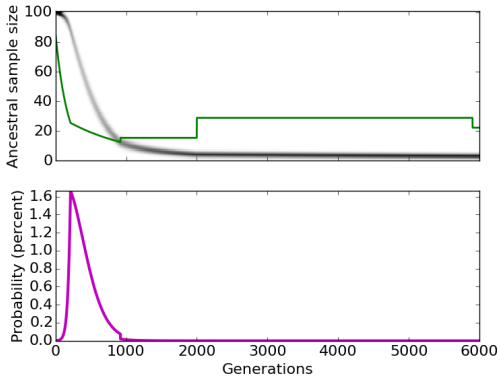
(a) Constant N.



(b) Double Bottleneck.

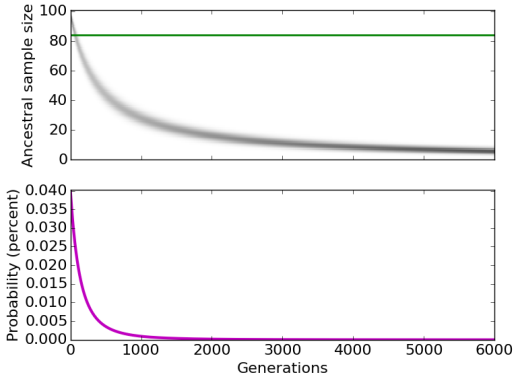


(c) Exp1.

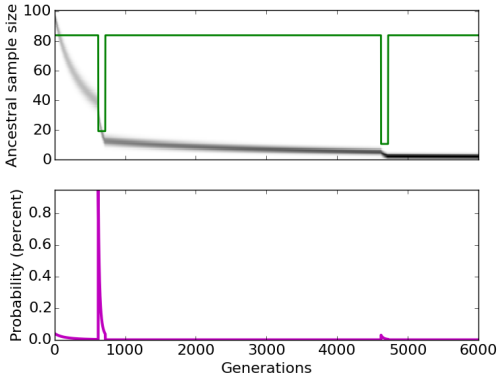


(d) Exp2.

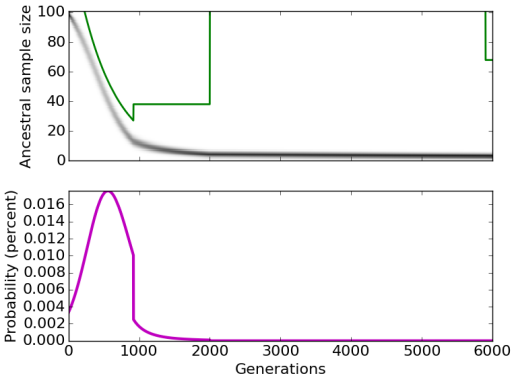
Figure 3: All four plots, which correspond to four different demographic models, are based on the algorithm to calculate the probability of coalescence allowing only single double collisions. In each case, the algorithm is applied with $n = 100$ at $t = 0$. Conditioned on no violation in any backward WF step from 0 to t (generations), there is a certain probability that the ancestral sample size at t is k. The upper panel is a heat-map of those probabilities, with black being 1 and white 0. The green line is a graph of $1.47 \times N(t)^{0.31}$. The lower panel is a graph of the probability of a violation in the backward WF step from generation t to $t+1$ conditioned on no violation from 0 to t.



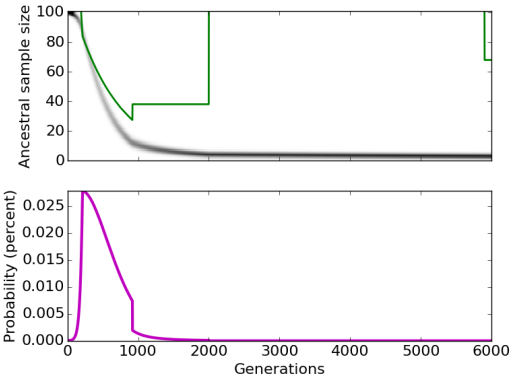
(a) Constant N.



(b) Double Bottleneck.



(c) Exp1.



(d) Exp2.

Figure 4: Same as Figure 3 except that the plots are based on the algorithm for calculating the probability of WF coalescence without triple collisions. The violations in this figure would be triple collisions, whereas the violations were simultaneous double collisions or triple collisions in the earlier figure. The green line in this figure graphs $0.92 \times N(t)^{0.49}$.

Our interpretation of the phenomena in Figure 3 (c) and (d) is as follows. In both cases, the heat-maps of ancestral sample sizes (conditioned on no violation) 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 of a violation appears to be located near the inflection point.

In Figure 4, the same phenomena are in evidence, with a violation here being a triple collision and not a simultaneous double collision or a triple collision as in Figure 3. Some of the phenomena are a little more prominent here. For example, a small spike in the conditional probability of a violation is visible even at the second bottleneck in part (b) of the figure.

Discussion

For a sample size of 60, the probability of deviation from the Kingman coalescent is higher than 50% in the human species. In the human species, the number of samples is often much greater than 60, and large sample effects cannot be ruled out.

What is the threshold beyond which large sample effects not captured by the Kingman coalescent are triggered? We have shown that the threshold is $N^{1/3}$ in an asymptotic sense and that the asymptotic theory is fully relevant to realistic population sizes.

However, $N^{1/3}$ is still a low threshold for the human species, and other studies (Bhaskar et al., 2014, Fu, 2006, Wakeley and Takahashi, 2003) have suggested that the sample frequency spectrum may not deviate too much for larger samples. We have shown that samples of size up to $N^{1/2}$ may experience simultaneous double collisions but not triple collisions in an asymptotic sense.

The sample frequency spectrum is only one of many possible statistics. Other statistics such as linkage disequilibrium may show deviations that are masked in the sample frequency spectrum. As noted by Kingman (1982b), the key to the coalescent is the probability distribution of the split of n samples between k ancestors in some past generation. Although this paper is developed from that point of view, the total variation distance between the probability distributions of the partitions under Kingman and WF is not yet known for large samples. Therefore, the question of how large the sample size must be to trigger effects that are not adequately handled by the Kingman model cannot be answered at the moment.

Nevertheless, there can be no doubt that the Kingman model produces unbiological effects for some large sample sizes, as we indicated in the introduction. Our computations show that even distant bottlenecks can impede the validity of the Kingman model. Non-African populations experienced bottlenecks during their first passage out of Africa and then again in later peregrinations around the globe. Whether unbiological effects of the Kingman model are to be accounted for in inferences of ancient bottlenecks is another interesting unanswered question.

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 the* $|A_j| = n_j$ *is the cardinality of* A_j*, the probability that the partition into* k *sets is* $\{A_1, \ldots, A_k\}$ *is equal to*

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

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 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* $\exp\left(-\frac{k^2}{2n}\right) \geq q(k,n) \geq \exp\left(-\frac{7k^2}{n}\right) \geq 1 - \frac{7k^2}{n}$.

Proof. The probability $q(k,n)$ is zero if $2k \geq n+1$ 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 $2k \leq n$. 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

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

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

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

for some $u \in [\frac{1}{2}, \frac{4}{5}]$. By similarly elementary arguments,

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

and

$$\frac{1}{m^2} + \frac{1}{(m+1)^2} + \cdots + \frac{1}{(n-1)^2} = \left(\frac{1}{m} - \frac{1}{n}\right) + u\left(\frac{1}{m^2} - \frac{1}{n^2}\right), \tag{3}$$

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

From the formula for p and (1), 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 $u \in [\frac{1}{2}, \frac{4}{5}]$. The application of (1) is justified because $3k \leq n$ implies $k/(n-k+1) < 1/2$. Applying (2) and (3), 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 $u_1 \in [0,1]$, $u_2 \in [\frac{1}{2}, \frac{3}{4}]$, and $u_3 \in [0, \frac{3}{4}]$.

Thus,

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

where the second inequality is obtained using (1). 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:

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

□

Lemma 3. *Consider the application of a single backward WF step to a sample of size n with parental population of size $2N$. Let p_d be the conditional probability that there is a single double collision given that there is a collision. Then*

$$p_d \geq 1-\frac{(n-2)(n-1)}{4N}.$$

Proof. Let A_{12} be the event that samples 1 and 2 collide (in one generation), more specifically 1 and 2 have the same Wright-Fisher parent. Obviously $\mathbb{P}(A_{12}) = \frac{1}{2N}$.

Let $A_{12}^{(t)}$ be the event that 1 and 2 collide and that one of the other $(n-2)$ samples has the same parent as 1 and 2, implying a triple collision 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 $\mathbb{P}(A_{12j}) = \frac{1}{(2N)^2}$, we have $\mathbb{P}(A_{12}^{(t)}) \leq \frac{(n-2)}{(2N)^2}$.

Let $A_{12}^{(d)}$ be the event that 1 and 2 collide and that there is some other pair that collides, implying two double collisions 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 $\mathbb{P}(A_{12,jk}) = \frac{1}{2N} \times \frac{2N-1}{(2N)} \times \frac{1}{2N} \leq \frac{1}{(2N)^2}$, we have $\mathbb{P}(A_{12}^{(d)}) \leq \binom{n-2}{2}\frac{1}{(2N)^2}$.

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

$$\begin{aligned}\mathbb{P}(\tilde{A}_{12}) &\geq \frac{1}{2N}-\frac{(n-2)}{(2N)^2}-\binom{n-2}{2}\frac{1}{(2N)^2}\\ &= \frac{1}{2N}\left(1-\frac{(n-2)(n-1)}{4N}\right).\end{aligned}$$

The probability that there is a single double collision 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 collision, $\mathcal{C} = \cup_{j,k} A_{jk}$, union over $1 \leq j < k \leq n$. Therefore, the probability of a collision $\mathbb{P}(\mathcal{C}) \leq \binom{n}{2}\frac{1}{2N}$. The lower bound for p_d in the lemma is obtained by simplifying $(2N)\mathbb{P}(\tilde{A}_{12})$. □

Theorem 4. *The WF genealogy of a sample of size $N^{1/3-\epsilon}$, $\epsilon > 0$, includes at most one double collision in any given generation, 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 double collision in some backward WF step before coalescing to a single ancestor. 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,

$$\begin{aligned}\mathbb{P}(\mathcal{D}_k) &= \mathbb{P}\left(\text{ancestral sample never of size } k-1 \middle| \text{sample of size } k\right) \\ &\quad \times \mathbb{P}\left(\text{ancestral sample is of size } k \text{ in some generation}\right) \\ &\leq \frac{(k-2)(k-1)}{4N}.\end{aligned}$$

Therefore,

$$\mathbb{P}(\mathcal{D}) \leq \sum_{k=3}^{n} \frac{(k-3)(k-1)}{4N} \leq \frac{n^3}{4N}.$$

If $n = N^{1/3-\epsilon}$, $\mathbb{P}(\mathcal{D}) \leq N^{-3\epsilon}/4$, which converges to zero as $N \to \infty$. In the complement of $\mathcal{D}$, WF coalescence follows the Kingman model proving the theorem. □

Let $\mathcal{D}_n$ be the event that there are more than two double collisions or a triple collision or worse. Then

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

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 collisions 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 double collisions. This bound simplifies to

$$\mathbb{P}\left(\mathcal{D}_n \middle| 1 \text{ and } 2 \text{ collide}\right) \leq \frac{n}{N} + \frac{n^4}{6(2N)^2}.$$

This is an almost correct bound for the conditional probability of $\mathcal{D}_n$ given *any* collision, 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 collide 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*

$$p_1 \leq \frac{n-k}{2N-j+1} \leq \frac{n}{2N-n}.$$

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 $2N-j+1$ 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 $\frac{1}{2N-j+1}$ 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 that parent is distinct from the parent of j and k. We have*

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

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 collision given $\mathcal{C}_{j,k}$ is

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

The first factor accounts for the first member of the triple having to choose a parent other than those of $[j]$ and $1/(2N-j+1)$ is the probability that the second or third member choose the same parent as the first, subject to $\mathcal{C}_{j,k}$. Therefore,

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

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*

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

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

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

Therefore,

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

as claimed in the lemma. □

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

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

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 double collisions in addition to the collision 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. *The WF genealogy of a sample of size $N^{\frac{c}{2c+1}-\epsilon}$, $\epsilon > 0$, includes at most c simultaneous double collisions and no triple collisions 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 collision or more than c double collisions. From the previous lemma,

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

Let $\mathcal{C}$ denote the event that a collision has occurred in a backward WF step with a sample size of ℓ. Evidently, $\mathcal{C}$ 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 collision in lexicographic order is between sample j and k. Therefore,

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

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

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

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. Let $q(n, 2N)$ denote the probability of a coalescence event in a single backward WF step. Then

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

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

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

The probability it is a coalescence is equal to

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

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

In particular, denote the event that the genealogy of a sample of size n (in either WF or Kingman model) includes exactly one mutation by $\mathcal{C}$. The probability that the mutation strikes when the WF ancestral sample size is k is equal to

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

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

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

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

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

Thus, μ_n is the probability that a mutation is the first event to strike a sample of size n conditioned on $\mathcal{C}$.

Let $f(j,n)$ be the probability that j out of n samples are mutants conditioned on $\mathcal{C}$. The recurrence for $f(j,n)$ is

$$f(j,n) = \mu_n[j=1] + (1-\mu_n)\left(f(j,n-1)\left(1-\frac{j}{n-1}\right) + f(j-1,n-1)\frac{j-1}{n-1}\right). \quad (4)$$

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 $q(j,2N) = j(j-1)/4N$ following the Kingman model. The exact solution of the recurrence

$$\tilde{f}(j,n) = \tilde{\mu}_n[j=1] + (1-\tilde{\mu}_n)\left(\tilde{f}(j,n-1)\left(1-\frac{j}{n-1}\right) + \tilde{f}(j-1,n-1)\frac{j-1}{n-1}\right). \quad (5)$$

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. $\frac{n(n-1)}{4N}\left(1-\frac{n^2}{4N}\right) \le q(n,2N) \le \frac{n(n-1)}{4N}$.

Proof. The proof of Lemma 3 shows that $\mathbb{P}(\tilde{A}_{12}) \ge \frac{1}{2N}\left(1-\frac{n^2}{4N}\right)$. The proof of the lower bound follows from $q(n,2N) \ge \frac{n(n-1)}{2}\mathbb{P}\left(\tilde{A}_{12}\right)$ and the upper bound is obviously true. □

Lemma 11. *For $n^2 \le 2N$,*

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

for $n^2 \le 2N$.

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 $q(j,2N) = \frac{j(j-1)}{4N}(1-s_j)$, then by the previous lemma $s_j \in [0, j^2/4N]$.

To obtain the lower bound in the lemma, set $s_j = 0$ in the denominator and $s_n = n^2/4N$ in the numerator.

To obtain the upper bound in the lemma, set $s_n = 0$ in the numerator and replace s_j by $\frac{n^2}{4N} \ge \frac{j^2}{4N} \ge s_j$ in the denominator. Finally, observe that $\left(1-\frac{n^2}{4N}\right)^{-1} \le \left(1+\frac{n^2}{2N}\right)$ if $n^2 \le 2N$. □

Lemma 12. *If $n \leq N^{1/3-\epsilon}$, 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 $|ab - \tilde{a}\tilde{b}| \leq |a-\tilde{a}|\,|b| + |b-\tilde{b}|\,|\tilde{a}|$. Subtracting (4) and (5), we get

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

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

Summing these inequalities, we have

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

Because $\tilde{\mu}_n < 1$ for $n > 2$ and $f(j,n) \equiv \tilde{f}(j,n)$ for $n = 2$, we have

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

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 double collision or a triple collision converges to zero as $N \to \infty$. Thus, we may invoke the previous lemma and infer this theorem. □

References

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

A. Bhaskar, A. G. Clark, and Y. S. Song. Distortion of genealogical properties when the sample is very large. *Proceedings of the National Academy of Sciences*, 111:2385–2390, 2014.

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

H. Chen, J. Hey, and K. Chen. Inferring very recent population growth rate from population-scale sequencing data: using a large-sample coalescent estimator. *Molecular Biology and Evolution*, 32:2996–3011, 2015.

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

R. Durrett. *Probability models for DNA sequence evolution*. Springer Science & Business Media, 2008.

W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112, 1972.

Y. Fu. Exact coalescent for the Wright-Fisher model. *Theoretical Population Biology*, 69: 385–394, 2006.

R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, NJ, 2nd edition, 1994.

S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, C. D. Bustamante, D. L. Altshuler, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108: 11983–11988, 2011.

R.C. Griffiths. Coalescent lineage distributions. *Advances in Applied Probability*, 38:405–429, 2006.

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

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

A. Keinan, J. C. Mullikin, N. Patterson, and D. Reich. Measurement of the human allele frequency spectrum demostrates greater genetic drift in east asians than in europeans. *Nature Genetics*, 39:1251, 2007.

J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19: 27–43, 1982a.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982b.

D.E. Knuth. *The Art of Computer Programming*, volume 1. Addison-Wesley, NJ, 3nd edition, 1997.

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

S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26:119–164, 1984.

J. Wakeley and T. Takahashi. Gene genealogies when the sample size exceeds the effective size of the population. *Molecular Biology and Evolution*, 20:208–213, 2003.

G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7:256–276, 1975.