Theoretical and Numerical Analyses of Deviations Between Kingman's Coalescent and the Wright-Fisher Model

by

Andrew Melfi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied and Interdisciplinary Mathematics)
in The University of Michigan
2019

Doctoral Committee:

Assistant Professor Jonathan Terhorst, Co-Chair
Professor Divakar Viswanath, Co-Chair
Professor Daniel Burns
Professor Daniel Forger
Professor Sebastian Zöllner

Andrew Melfi
melfi@umich.edu
ORCID ID: 0000-0003-3140-5759

To my mom and dad.

ACKNOWLEDGEMENTS

I owe many people a great deal of thanks for getting me to where I am today. First, I would like to thank my father, and my high school math teachers, Cecilia Anderson and Pavel Sikorskii, for nurturing my skills and helping me to discover math as an enjoyable and interesting subject. I would like to thank my undergraduate mentor Andrew Christlieb, for allowing me to take a first step into the world of research, and for convincing me to continue my studies at the University of Michigan.

I would like to thank the Mathematics Department at the University of Michigan for allowing me to pursue my studies in applied mathematics. The courses I've taken have been fantastic, taught by helpful instructors, too many to name. I'm grateful for the opportunity to teach: I believe I've learned as much from that experience as I've taught my students.

I owe my mathematics advisor Divakar Viswanath a great deal of thanks for the wisdom he's bestowed and the patience he's shown me over the years. Together we entered a new field, and the resulting adventure is one I will always look back on fondly. Without his help I would be nowhere near the mathematician I am today. Additionally, I would like to thank my co-advisor Jonathan Terhorst for his help in understanding what it means to work in population genetics. Also, let me thank my committee members: Daniel Forger, Daniel Burns and Sebastian Zöllner.

Thank you to all the friends I've made in Ann Arbor. Thanks to you, my time here has been constantly enjoyable, and passed faster than it feels like it should

have. Particularly let me thank my once roommate Michael Newman, who offered nearly enough advice to offset the distractions he's introduced to my life, and my climbing/hiking friends Jeremy Hoskins and Vivienne Baldassare, who've joined me on several adventures over the years.

Finally, I would like to thank my mom and my dad, for raising me with a sense of curiosity that has made all my academic pursuits enjoyable. Over my time here, they've constantly listened to my concerns and complaints and offered me encouragement and assistance whenever I've needed it.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

The Kingman Coalescent is a commonly used model in genetics, which is often justified with reference to the Wright-Fisher (WF) model. In this thesis we seek to attain a deeper understanding of the relationship between these two models, particularly by quantifying under what conditions the models are similar, and by understanding the ramifications of deviations between the models outside those conditions.

In Chapter 2, we investigate one source of deviation between the two models, that they have different partition distributions. We find an asymptotic bound on sample size relative to effective population size under which the partition distributions are identical. We additionally find similar asymptotic bounds under which no triple mergers will occur in the Wright-Fisher model. Furthermore, we use numerical methods to show that these bounds are generally applicable at finite sample and population sizes.

In Chapter 3, we investigate the deviation between the site frequency spectrum (SFS) under the WF model and the coalescent model. There are two sources of this deviation. One is that there is a mismatch in rates of merger between the two models. The other is the aforementioned difference in partition distributions. The mismatch in rates raises the probability of singletons under WF, but the difference in partition distributions lowers it. These two effects are opposing everywhere except at the tail of the frequency spectrum. The WF frequency spectrum only begins to significantly depart from that of the coalescent at sample sizes close to the population size. We

examine the case where the sample size is assumed to be equal to the population size N and find the total variation distance between WF and coalescent to be only 1 % for populations of size 20000. Therefore we conclude that the coalescent is a good approximation for WF for the site frequency spectrum of large samples.

In Chapter 4, we introduce an algorithm which allows us to generate the SFS under the coalescent with a time-varying population size and mutation rate. Using this algorithm we explore the effects of a variable mutation rate on the SFS. We find that the SFS changes substantially as a result of varying mutation rates even for small samples.

CHAPTER 1

Introduction

1.1 The sequencing revolution

In 2001, the Human Genome Project released, at a cost of approximately 3 billion dollars, the full sequence of DNA in the human genome [40]. Since then, the costs associated with sequencing have fallen exponentially, and today it costs less than $1000 to fully sequence a particular human genome [80]. With relatively inexpensive sequencing available, researchers are able to use this data to make inferences about which parts of the genome are responsible for traits (phenotypes) such as height, obesity, or genetic diseases. The primary tool used to make these inferences is called a Genome Wide Association Study (GWAS), where individuals are grouped according to the phenotype of interest, and statistical analysis is used to find genetic variations which may explain the differences between groups.

While establishing links between genes and traits is an interesting medical application of sequencing data, this only scratches the surface of what is possible. The modern genetic landscape is the result of a complex genealogical history, so by modeling it is possible to make inferences about ancient events and selective forces which led to who we are today. For example, some research sequences the DNA of ancient hominids such as Neanderthals or Denisovans and seeks to understand their relationship with modern humans [69, 77]. Other research compares the DNA of

groups of modern humans seeking to determine dates of admixture, divergence or bottleneck events [66, 35, 54, 72]. Another avenue of research seeks to find regions in the genome affected by natural selection [71], such as the proliferation of the lactase enzyme in European populations. All of this research is dependent on proper modeling of genealogies.

We will discuss two of the most common models of genetic history, the Wright-Fisher model and the Kingman coalescent, but first we will spend a bit of time discussing the basics of the human genome and how DNA is passed on within a single generation.

1.2 Basics of the human genome

A single human's DNA is made up of 23 pairs of chromosomes and a small DNA molecule within mitochondria. The 23rd pair of chromosomes are those which determine sex and are called allosomes. The first 22 pairs are known as autosomes. Each autosome in a pair encodes the same set of genes, and one comes solely from the father and one solely from the mother.

A single chromosome consists of a DNA molecule which consists of two strands of nucleotides. Each nucleotide contains one nucleobase (cytosine (C), guanine (G), adenine (A) or thymine (T)) and it is the ordering of these bases which encodes genetic information. The two strands are complementary, with the second strand having bases entirely determined by the first. A pairs with T and C pairs with G. So if the first strand read ACT, the other strand would have the pattern TGA. One strand contains the bases as they appear in messenger RNA, this is known as the sense strand, and the other is referred to as the antisense strand. So we can describe a chromosome in a person by writing a long list of nucleotides.

The largest chromosomes within a human consist of approximately 250 million base pairs, and the smallest approximately 50 million base pairs. So in a sense, to fully describe the DNA in a person we need to record 3.3 billion bases, which would be nearly a gigabyte of information. However it is not quite that bad. The majority of the human genome (99.4%) [1] is exactly the same between two people, so all we need to describe a human genome is to list at what locations variations occur. Even more conveniently, most variation between individuals are single nucleotide polymorphisms (SNPs pronounced 'snips') meaning a single nucleotide differs but the surrounding nucleotides are the same. The rate of mutation in humans is quite low relative to the size of the genome, so these polymorphisms typically only have two forms. Therefore, when comparing a group of people, we can classify an individual using a single bit for each SNP, typically 0 for the more common form, and 1 for the less common.

Humans are known as diploid organisms, meaning that they have pairs of each chromosome, one from each parent. Our models will assume that individuals are haploid (meaning having just one of each chromosome) and each child has a single parent. This reduces the complexity by a great deal and makes modeling much easier. Why would such a simplified model yield any useful information about a diploid species such as humans? To understand this, we will briefly look at the process by which humans create gametes (which pass genetic information to their children), meiosis.

From the perspective of a pair of autosomes, the process of meiosis takes a single diploid cell, exchanges DNA between those homologous autosomes through a process called recombination, and produces (after an initial stage of copying) four haploid gametes containing one mixed autosome each. Two gametes, one from each parent,

merge to form the diploid set of chromosomes in a child.

If we restrict ourselves to working with one site on a chromosome, we can ignore the mixing of recombination and meiosis becomes the following process: Two of the four gametes resulting from meiosis receive a copy of the site from the first paired autosome, and the other two gametes receive a copy of the site from the other paired autosome. Therefore, if we consider a human as two haploid individuals we can consider human reproduction as the process by which one haploid individual from each parent successfully copies itself to produce a new haploid individual. Some small artifacts are created compared with pure asexual haploid reproduction such as the inability for one chromosome to completely dominate the next generation, but in random models this type of event is so unlikely as to be negligible.

Of course, by ignoring the process of recombination we are losing information regarding associations between nearby sites on a chromosome (linkage). In some contexts, recombination is a useful construct to investigate. For example, attempts to use DNA to identify ancient human admixture events rely on artifacts produced by recombination [35, 54, 72].

Additionally, allosomes and mitochondrial DNA reproduce in a way distinct from the autosomes. Particularly the Y chromosome undergoes no recombination and is of course only inherited by males, the X chromosome undergoes less recombination than the autosomes (only in female meiosis), and mitochondrial DNA is only inherited matrilineally. Because of these differences we typically focus on autosomal DNA, though again quite a bit of research is done which takes advantage of the idiosyncratic inheritance of non-autosomal DNA.

1.3 The Wright-Fisher model

Consider a population of N haploid individuals with two alleles, A and B at a given locus. The Wright-Fisher (WF) model assumes all of the individuals in the population die each generation and are replaced by their offspring. The population size N is assumed to be constant over time. Additionally, we assume that neither allele confers additional fitness on the individual.

Let $K(t)$ represent the number of copies of the A allele in generation t with the original population corresponding to $t = 0$. Suppose $K(0) = i$. Then $N - i$ is the number of copies of the B allele. The frequency of A in this generation is then $p = i/N$, and the frequency of B is $1-p$. Under the WF model, each offspring draws uniformly at random with replacement from the previous generation and obtains the allele from that parent. Therefore,

$$P_{ij} = \binom{N}{j} p^j (1-p)^{N-j} \tag{1.1}$$

is the probability that an allele with i copies in the current generation is found with j copies in the next generation. The expectation of the number of copies of A will be $K(0)$ for any given generation, but over time K will drift randomly according to the Markov chain with transition probabilities P_{ij}. Eventually, either K will reach 0 (the extinction of the A allele) or reach N (the fixation of the A allele).

To understand the reason why (1.1) might make biological sense, consider the following scenario: Before dying, each individual in the population produces a large number of gametes, each of which might become individuals in the following generation. While there are a large number of potential offspring, the population size is tightly controlled so only N of these may make it into the next generation. The proportion of gametes containing the allele A is i/N, and because the alleles provide

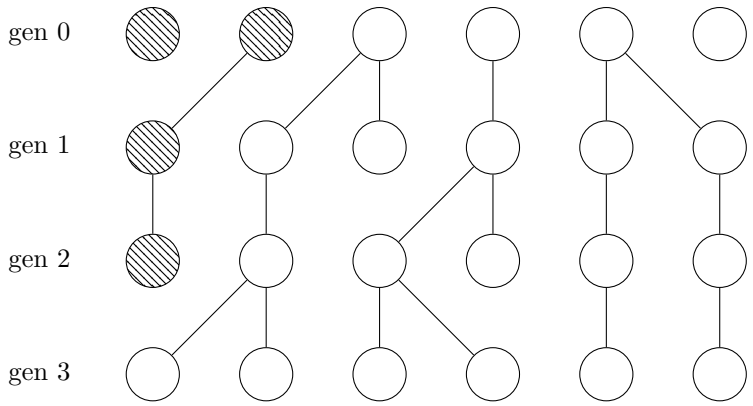


Figure 1.1: In the above diagram, individuals with allele A are represented by white circles and individuals with allele B are represented by shaded circles. Lines connect parents and children and the model evolves according to the Wright Fisher model starting with generation 0. After 3 generations, the B allele has died out so we say the A allele is *fixed* within the population.

equal fitness, all gametes have equal probability of being selected as the N individuals in the next generation. This selection can be thought of as N trials, each with an i/N probability of success (containing the A allele). Because the pool of gametes is so large, the sampling does not deplete it, so the probability does not change between trials. This is exactly the setup which produces the binomial distribution.

The WF model was conceived in the 1930s [18, 95], before there was knowledge of the structure of DNA. It was originally used to study the effects of genetic drift, the process by which randomness in reproduction alters the frequencies of genetic variants. (cit)

Of course this basic model may be augmented to capture other effects which occur in nature. For example, we could do away with the assumption that the alleles both provide equal fitness to study the effects of natural selection. Typically this is modeled by an adjustment to the frequencies of individuals in the large pool of gametes described above[89]. As another example, we could allow for the possibility of mutations which would introduce new alleles into the population.

While this model provides a simple framework for how genes may propagate into

future generations, it does not lead us to any answers to our initial questions regarding the origins of genetic variation within a modern population. To approach this we need to make a few modifications.

1.4 Reverse Wright-Fisher

Suppose we have a sample of 10 haploid individuals who reproduce according to the WF model. Again consider a single SNP with alleles A and B. If we go back in time one generation, it is possible that two of these individuals have the same parent. This would imply that the siblings had the same allele. We say that their lineages coalesced. Now, continuing to move back in time following our 9 lineages, eventually there would be another coalescence event. We would be reduced to 8 lineages. Go back far enough and eventually our 10 lineages will coalesce into a single one. We call this the most recent common ancestor.

This process may be applied to human populations as well. If we consider a single SNP between a group of individuals, going back far enough in time we will find that SNP to be inherited from some distant common ancestor. This would seem to imply that there should be no genetic variation, as coalescence requires that the children have the same allele. In fact, all genetic variation is due to a mutation somewhere in the genealogy between the most recent common ancestor and the modern sample.

To account for this in the WF model we will make two modifications. First, we will reverse the direction of time. We will start with a modern group we refer to as a sample, and track its lineage going backwards in time. Second, we will introduce a probability of mutation. In every generation each child will have some probability of having a mutation (neutral mutation). In this case, all of its descendants will have an allele differing from the rest of the genealogy[1].

[1] Under the infinite alleles model

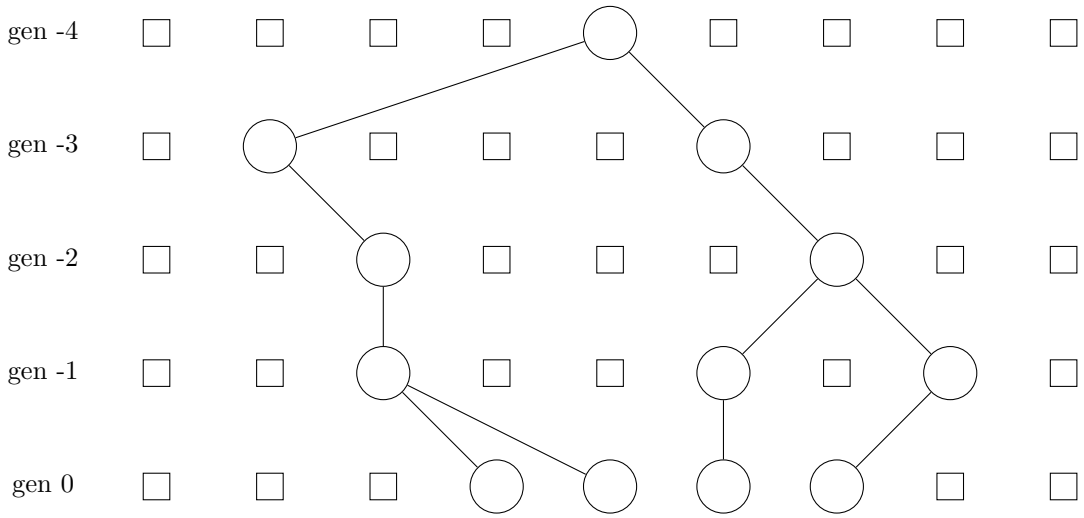
Typically we will assume the mutation rate is quite low and that exactly one mutation will occur throughout the entire genealogy. This is a reasonable assumption to make because as mentioned earlier, the majority of the human genome is identical between individuals, implying that coalescence occurred before a single mutation. We are interested in locations where a variant allele exists within our sample, so we have preselected for only locations where a mutation has occurred.

Now let us formalize the reverse WF model. Consider a sample of size n within a population[2] of size N with a mutation rate of μ per individual per generation. Each generation we will go back in time and perform the following two steps. First, each remaining lineage will determine whether a mutation has occurred. Second, each lineage will select an integer between 1 and N uniformly at random. If any two lineages select the same number it implies they have the same parent, so a coalescence occurs, in which case for subsequent generations we will only track $n-1$ lineages. We refer to the representative of each remaining lineage as the ancestral sample. This process is repeated until the ancestral sample has size 1, meaning we have found the most recent common ancestor of the initial sample.

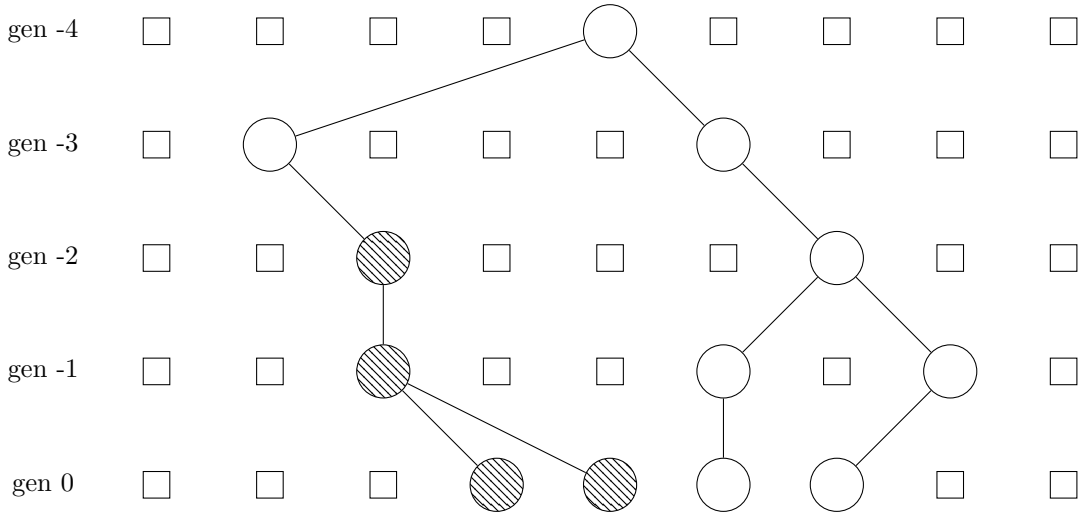
Under the assumption that the population size N is constant it is possible to create a tractable algorithm which calculates the probability of $m \in \{1, \ldots, n-1\}$ mutant alleles in the modern sample, known as the sample frequency spectrum (SFS). We describe such an algorithm in Chapter 2.

Each generation under the reverse WF model, there is a $1/N$ chance that 2 specific lineages will coalesce. Therefore, the expected time for a sample of size 2 to coalesce is N generations. We can also note this means the expected time for k lineages to coalesce into $k-1$ is $N/\binom{k}{2}$ generations.

[2]Here, a sample refers to the modern set of individuals of interest, and the population refers to the entire breeding population.



(a) The above figure represents the reverse Wright Fisher model applied to a sample of size 4 within a population of constant size 9. The boxes represent members of the population who are not part of the ancestry of our sample. In this case, within 4 generations our sample has coalesced to a common ancestor.



(b) Here we see what happens if a mutation strikes 2 generations before the present, under the same genealogy. The mutated allele is represented by a shaded circle. All descendants of this individual have inherited the mutation.

Figure 1.2: Reverse Wright-Fisher with and without mutations.

A natural way to reduce the computational complexity of the reverse WF coalescent would be to replace it by a continuous time model, doing away with the concept of discrete generations. This is the idea which inspires Kingman's coalescent which we will describe presently.

1.5 Kingman's Coalescent

With the objective of producing a model more suitable for analysis than the reverse WF model, Kingman noted that the expected time to most recent common ancestor under WF for a sample of n individuals is proportional to N[52, 51]. Therefore, he thought a suitable scaling for time would not simply be generations, but units of N generations. Under this assumption he allowed N to scale to infinity and obtained a model with several desirable properties now known as the Kingman coalescent or often simply the coalescent.

While Kingman uses a scaling of N generations per unit of time, it is convenient to still think in terms of generations. If we let T_k represent the time required for k lineages to coalesce into $k-1$ (in generations), under the coalescence model the coalescence times T_i, $i \in \{2, \ldots, n\}$ are independent and exponentially distributed as [89]

$$f_{T_i} = \frac{1}{N}\binom{i}{2} e^{-\frac{1}{N}\binom{i}{2}}. \tag{1.2}$$

Whenever a coalescence event occurs, two lineages are selected uniformly at random, and they coalesce.

To incorporate mutations into this continuous model, we note that under the reverse WF model, the number of mutations per generation is given by $k\mu$, where k is the number of lineages remaining in that generation. This implies that the expected time for a mutation to occur with k lineages is $1/(k\mu)$. So while there are k

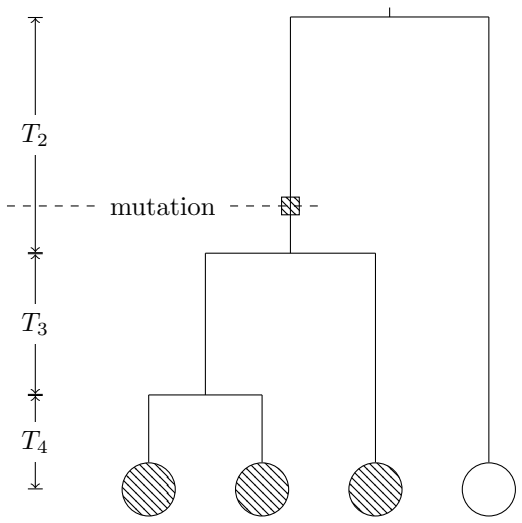


Figure 1.3: Here we see the Kingman coalescent run with a sample size of 4. After T_4 generations, 2 lineages are chosen to coalesce. Again after an additional T_3 generations. However, before they fully coalesce to a common ancestor, a mutation event occurs. This mutation is inherited by all descendants of this lineage.

lineages remaining we model the mutation event by an exponential distribution with parameter $k\mu$. If a mutation occurs before a coalescence event reduces the number of lineages, we select a lineage at random and choose it to be hit by a mutation.

This continuous time model is more convenient analytically and computationally than the WF model. For example, returning to the SFS, it is possible to derive analytically the SFS for constant population size conditioned on exactly one mutant:

$$p(j,n) = \frac{1/j}{\sum_{k=1}^{n-1} 1/k},$$

where $p(j,n)$ represents the probability of j mutant alleles given a sample size of n [11, 29].

1.6 Differences between Wright Fisher and coalescent

The Kingman coalescent and Wright Fisher model are two different, widely-used models within population genetics. Researchers have long accepted that for sample sizes n much smaller than N, the coalescent approximates WF well [20]. Unfortunately, what "much smaller" means has not been analyzed rigorously. We seek to

make clear under exactly what conditions the two models are close, and exactly what "close" means. We tackle this in two different ways.

One of the main sources of deviation between the two models is that the models generate partitions of the sample whose probability distributions are different. For example, if 10 samples in a single generation are known to have one of two parents, under the WF model, the split of the 10 children between the parents has a binomial distribution. However, under the coalescent the split has a uniform distribution. This difference is because under the coalescent, only two lineages can merge at a time as opposed to WF, where in a single generation multiple samples may have the same parent. In Chapter 2, we find an asymptotic bound on n which guarantees identical partition distributions. Furthermore, we use numerical simulations to show that this bound is applicable even for finite values of N.

The condition that the partition distributions must be identical is almost certainly too strict to use generally as a restriction on n. One of the most commonly used summary statistics generated by both models is the SFS. In Chapter 3, we seek to investigate for larger values of n (up to $n = N$) how large the deviation in SFS is between models. Additionally, we refine our asymptotics from Chapter 2.

1.7 Varying mutation rates

One implicit assumption in the models described above is that the mutation rate is constant. It has been known for quite some time that mutation rates differ across the genome (specifically CpG junctions are known to have higher mutation rates [12]). However, recently it has been found that various heptanucleotide contexts have up to 650 fold different mutation rates than others [3, 7]. This is of great importance with regards to using global minor allele frequency to generate a sample SFS.

Additionally, research shows that differing ages of reproduction results in a different per generation mutation rate [65, 76]. Evidence also shows that global (whole genome) mutation rates have been slowing down over the past million years [79].

All of this suggests a need to incorporate a varying mutation rate into genealogical models. In Chapter 4 we demonstrate an algorithm which efficiently calculates the SFS under the Kingman coalescent with a variable mutation rate. We additionally investigate a similar algorithm under the WF model.

CHAPTER 2

Single and Simultaneuous Binary Mergers in Wright-Fisher Genealogies[1]

2.1 Introduction

The Kingman coalescent [52, 51] 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 parts. For convenience, we will refer to lineages present in earlier generations as ancestral samples[2].

One of Kingman's motivations in deriving the coalescent [52, 51, 50] was to gain an understanding of the structure of Ewens' sampling formula [14, 11]. The coalescent gives an almost instantaneous derivation of Ewens' 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 [11]. In the backward in time genealogical process, the k sam-

[1] A modified version of this chapter, under the same title, is published in *Theoretical Population Biology* [58]

[2] The "ancestral sample" nomenclature is more intuitive for our purposes. However, in the context of the coalescent, the same concept is sometimes referred to as "lineage" or "ancestral lineage" [27, 29, 84].

ples 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 turns 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 [91]. Genealogies in WF as well as other exchangable models have been proven to converge to the Kingman coalescent [51, 62, 63]. 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 $N = 2 \times 10^4$ [44, 82]. Thus, there is a need to advance convergence theory beyond the assumption of constant sample size. The beginnings of such a convergence theory is presented in this chapter by considering the genealogical coalescence process using Kingman's model as well as the WF model.

2.1.1 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 [52, 51]. 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 comparable to N, there will be simultaneous binary mergers as well as triple mergers in a single WF generation [4, 20]. 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 $n(n-1)/N$, it may still be able to capture the multiple mergers that occur in a single WF generation [5, 20].

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 [20, 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

$$s_{WF}\left(\frac{n}{N}\right) = s_K\left(\frac{n}{N'}\right), \tag{2.1}$$

where s_{WF} and s_K are functions depicted in Figure 3 of [20].

The other 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

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

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 [11, page 13, Theorem 1.6][29]. Thus, it is clear that although the Kingman coalescent can produce simultaneous binary mergers as well as triple mergers over a time interval corresponding to a single generation, the partitions it produces will have a different distribution from that of WF.

2.1.2 Convergence theory for sample sizes that increase with N

As implied by the classic birthday problem and its variants [4], 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 probability of $1-e^{-1/6}$ in the limit of large N. For sample sizes in-between, there will be simultaneous binary mergers (between distinct pairs of samples) in a single generation with high probability. By our convention, quadruple and higher mergers also count as triple mergers. Additionally, we will refer to any generation involving more than a single binary merger (simultaneous binary or triple) as a generation containing multiple 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 mergers [5, 10]. We prove a result (Corollary 2.2) 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 probabilty 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 available at `github.com/melfiand/lsample`.

2.1.3 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 $k = 1, \ldots, n-1$ [11, 29]. 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 who 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 [25, 46, 86]. Because of its pertinence to applications, the departure of the WF sample frequency spectrum from that of the coalescent has attracted attention. [92] 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), [92] 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 [5].

[20] derived an exact coalescent for WF. Like [92], he found that 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 [20] used computer simulations of the exact WF coalescent to study the sample frequency spectrum, [5] derived exact recurrences for the sample frequency spectrum as well as the expected number of triple mergers and other genealogical quantities. The algorithms of [5] 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. [84] obtained formulas for the size of the ancestral sample (number of lineages) as a function of the ancestral generation, assuming fixed population size. [29] 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, [27] obtained asymptotic approximations that are better numerically for large samples. Other authors, [8, 74, 75] have extended this work to handle coalescence and inter-coalescence times. In particular, [9] have observed that the number of segregating sites, an important statistic introduced by [94] and which marked the shift from infinite alleles to the infinite sites model [11], 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 [84] as shown by [5].

2.2 Results

The coalescent consists of two independent stochastic processes [51]. 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 $\{A_1, \ldots, A_k\}$ 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 [51]. 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 resulting in a reduction of the number of ancestral samples by 1.

The other part of the coalescent is the so-called death process [51], 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 with 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 process does not play any role in the convergence of the Kingman partition distribution. The death process governs the rates of coalescence, which can be adjusted independently,

as shown in (2.1).

The following theorem of [51] characterizes the jump chain completely via the Kingman partition distribution and does not depend upon the death chain:

Theorem 2.1. *[51] 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*

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

A conclusion we may draw from this is that if coalescence under the WF model consists solely of binary mergers, the resulting partition distribution is the same as that of the Kingman coalescent.

All theorems and corollaries stated in this section will be proved in Section 2.5. For the above theorem, we give a combinatorial proof of the Kingman partition distribution in the spirit of [28]. Kingman's proof is recursive [51, 11].

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

Corollary 2.2. *Suppose the set* $\{\{1\}, \ldots, \{n\}\}$ *undergoes* k *coalescences resulting in a partition of* $[n]$ *into* $n-k$ *sets. The probability 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* $k \geq 2$*, then additionally*

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

In this corollary, the falling power $n(n-1)\dots(n-k+1)$ is denoted $n^{\underline{k}}$ as recommended by Knuth [24, 53]. 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 not only look at bounds on n in terms of the population size N that allow only single binary mergers (with high probability), but also investigate bounds that allow simultaneous binary mergers.

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

Theorem 2.5. *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 $N \to \infty$.*

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 rate 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 corollary 2.2, 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 2.7. *Each backward WF step in the genealogy of a sample of size $N^{\frac{2}{5}-\epsilon}$,*

$\epsilon > 0$, *consists of zero, one or two binary mergers with probability converging to* 1 *in the limit of 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 partition distribution of the WF coalescent of samples of size $N^{2/5-\epsilon}$ or less is a mixture of that 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 2.7 (General Case)**.** *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.

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 multiple mergers in backward WF steps are simultaneous binary mergers.

While the above theorems show the equivalence of the partition distribution (or near equivalence) between WF and Kingman for various sample sizes, that does not immediately imply that commonly used statistics such as the sample frequency spec-

trum will be equivalent as well. The following theorem will show that the WF sample frequency spectrum converges to that of the coalescent for sample sizes smaller than $N^{1/3-\epsilon}$.

Let $\tilde{f}(k,n)$ 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 $\tilde{f}(k,n)=\frac{1/k}{\mathcal{H}_{n-1}}$ in the limit of zero mutation rate.

Theorem 2.12. *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.

2.3 Verification and visualization

In order to investigate the utility of our asymptotic theory when working with population sizes that are finite and that potentially vary with time, we use an algorithm to compute exact probabilities of merger event types each generation. More specifically, given $N(t)$, a function of effective population size t generations ago, we calculate $\phi_n(k,t)$, the probability that under the Wright Fisher model a sample of size n will coalesce into an ancestral sample size of k in generation t without any generations containing multiple mergers. Since the sample will never converge into an ancestral sample size of 0, we give $\phi_n(0,t)$ a special meaning: the probability that a multiple merger has occurred between generation 0 and t. Note that this may be

calculated in the following way:

$$\phi_n(0,t) = 1 - \sum_{k=1}^{n} \phi_n(k,t).$$

Additionally, the probability of a multiple merger between generations t and $t+1$ conditioned on at most single binary mergers in each step from 0 to t is

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

This formula will be used to visualize the effect of bottlenecks.

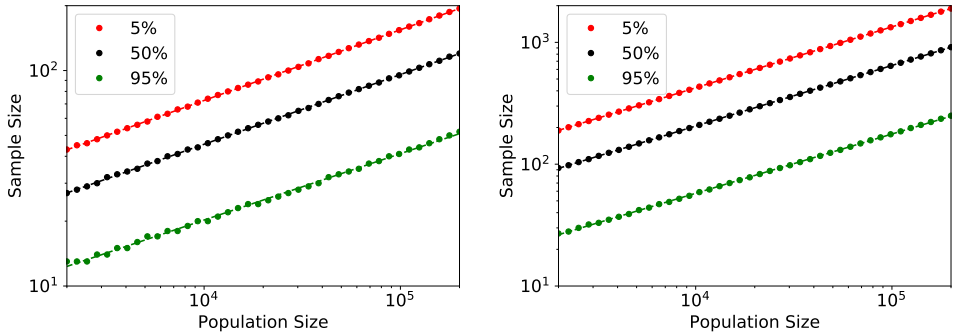
In order to calculate the probability that a sample of size n will contain any multiple mergers within its WF coalescence tree, we simply run the algorithm until $t = T$ where $\phi_n(1,T) + \phi_n(0,T) > 1 - \epsilon$, for some sufficiently small ϵ. We typically take $\epsilon = 10^{-4}$. Then $\phi_n(0,T)$ is the value of interest.

Using a small modification to our algorithm we additionally calculate $\psi_n(k,t)$, the probability that a sample of size n will coalesce to an ancestral sample size of k in generation t without any generations containing triple mergers. We can work with this the same way as $\phi_n(k,t)$. Both of these algorithms will be described in greater detail in Section 2.6

One may wonder how well the asymptotic theory from Theorems 2.5 and 2.7 applies to finite constant population size. To investigate this, we calculate $\phi_n(0,T)$ and $\psi_n(0,T)$ for various constant population size N and determine for each the cutoff population size n for which the probability of no multiple mergers or no triple mergers is 5%, 50%, or 95%. The results of these calculations are shown in Figure 2.1.

Sample sizes for which probabilities of coalescence with no multiple mergers are 5%, 50%, and 95% may be fitted as

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



(a) Only single binary mergers allowed.

(b) No triple mergers allowed.

Figure 2.1: 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.

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.

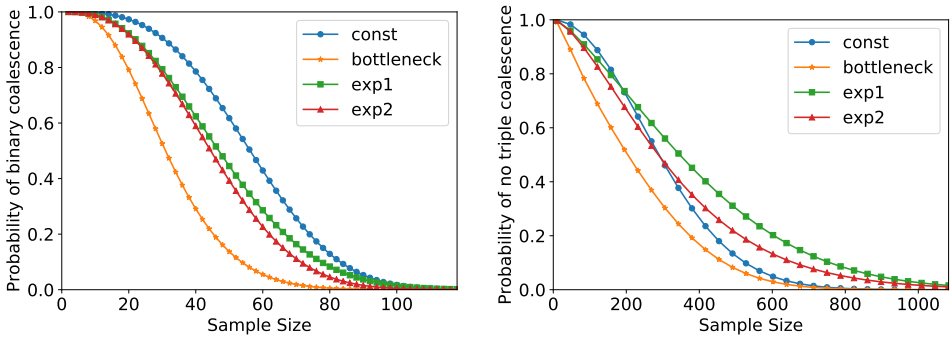
The fits for the no triple merger case are in even better agreement with the asymptotic theory. In this case, the sample sizes for which the probabilities of no triple mergers are 5%, 50%, and 95% are

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

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 approximately a factor of six.

Both ϕ_n and ψ_n allow for variable population sizes. The four demographic models of human population we consider are the same as in [5]. These models are:

- Constant population with $N = 2 \times 10^4$, which is the baseline assumption in human genetics [11].
- Constant population with $N(t) = 2 \times 10^4$ except for two bottlenecks: the first being $620 < t \leq 720$ with $N(t) = 1000$ and the second being $4620 < t \leq 4720$



(a) Only single binary mergers allowed

(b) No triple mergers allowed

Figure 2.2: 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.

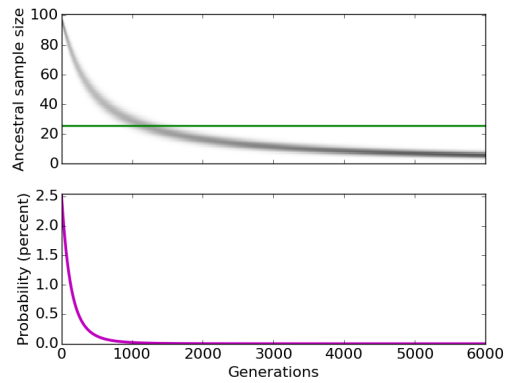
with $N(t) = 300$, a dropoff of nearly a factor of 100. This model is based on [46]

- 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 $N(t) = 3 \times 10^4$ for $2000 < t \leq 5900$, and $N(t) = 1.3 \times 10^4$ for $t > 5900$. This model is based on [25]. This model features a single exponential and is labeled `exp1` in Figure 2.2.

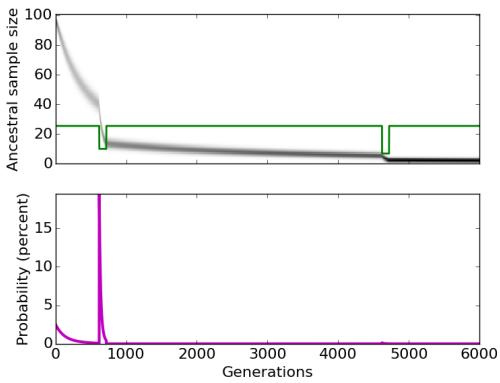
- Exponential decay for $0 \leq t \leq 214$ from $N(0) = 10^6$ to $N(214) = 2 \times 10^4$, exponential decay for $214 < t \leq 920$ with $N(920) = 2050$, $N(t) = 4000$ for $920 < t \leq 2000$, $N(t) = 3 \times 10^4$ for $2000t \leq 5900$, and $N(t) = 1.3 \times 10^4$ for $t > 5900$. This model is based on [86] This model features two exponentials and is therefore labeled `exp2` in Figure 2.2.

Figure 2.2 shows that the probabilities of multiple mergers and triple mergers in WF genealogies increase noticeably because of bottlenecks.

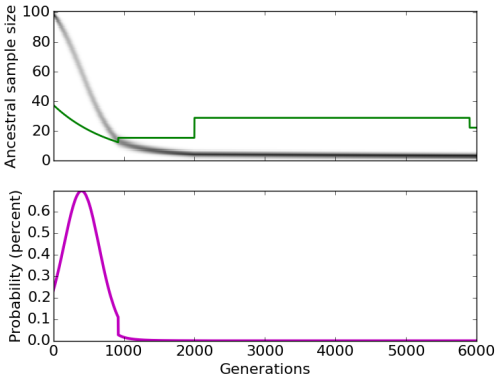
Figures 2.3 and 2.4 give a more explicit visualization of the effect of bottlenecks. In 2.3 (b), the distribution of possible ancestral sample sizes, conditioned on at most



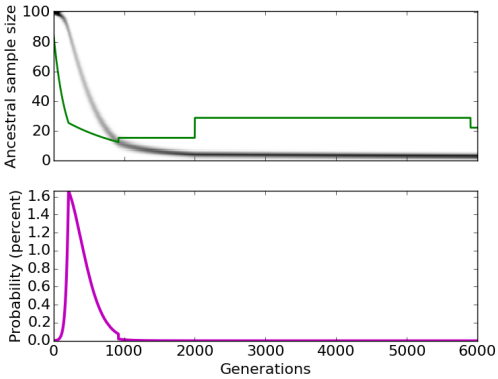
(a) Constant N.



(b) Double Bottleneck.

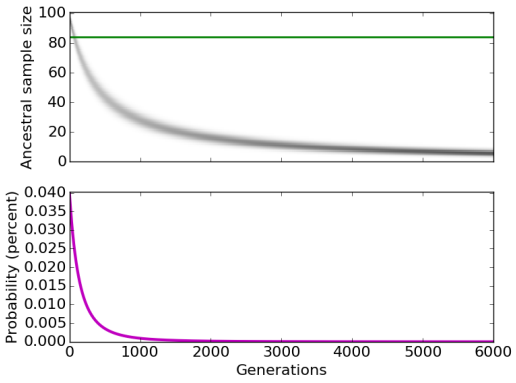


(c) Exp1.

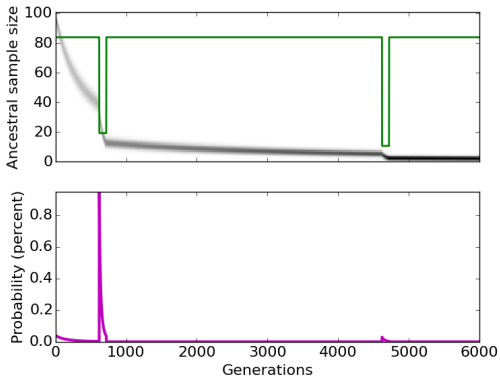


(d) Exp2.

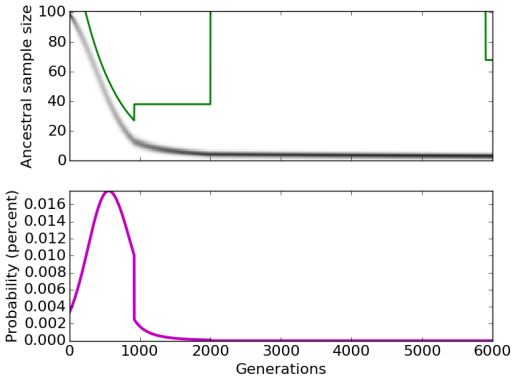
Figure 2.3: The upper panels in (a) through (d) are heat-maps of probabilities $\phi_n(k,t)$, with black being 1 and white 0. The green line is a graph of $1.19 \times N(t)^{0.31}$. The lower panels in (a) through (d) graph the conditional probability of a multiple merger per generation given no multiple mergers to that point. The plots (a) through (d) correspond to four different demographic models. The sample size is $n = 100$ in all plots.



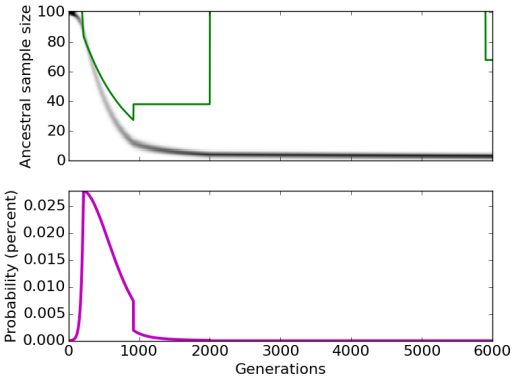
(a) Constant N.



(b) Double Bottleneck.



(c) Exp1.



(d) Exp2.

Figure 2.4: 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 of a triple merger per generation given no triple mergers to that point. As before, (a) through (d) correspond to four different demographic models with sample size $n = 100$.

a single binary merger in prior generations, noticeably shifts downwards when the first bottleneck is encountered. The conditional probability of a multiple merger also spikes at the first bottleneck. At the second bottleneck, there is no such prominent spike. However, the distribution of possible ancestral sample sizes, allowing no multiple mergers, noticeably shifts downward even at the second bottleneck, even though the bottleneck is more than 4500 ancestral generations back and the sample size is only 100.

Our interpretation of the phenomena in Figure 2.3 (c) and (d) are 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 conditional probability of a multiple merger per generation given no previous multiple mergers appears to spike near that inflection point.

In Figure 2.4, the same phenomena are in evidence. In fact they may be a little more prominent here. For example, a small spike in the conditional probability of a triple merger per generation given no previous triple mergers is visible even at the second bottleneck in part (b) of the figure.

2.4 Discussion

The roots of the Kingman coalescent may be found in the work of [14] and [94]. It was derived [52, 51] 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 [44, 82]. 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 2.1). 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.

2.5 Theorem proofs

Theorem 2.1. *[51] 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*

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

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 the leaves being $\{1\}, \ldots, \{n\}$. If disjoint subsets S_1 and S_2 coalesce, $S_1 \cup S_2$ occurs

as a vertex with S_1 and S_2 as its two children. Coalescences deeper in 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.

First, note that the number of coalescent trees with root $\{1, \ldots, n\}$ is

$$\frac{n!(n-1)!}{2^{n-1}}. \tag{2.2}$$

This is because the first union is any of $n(n-1)/2$ possibilities, the second any of $(n-1)(n-2)/2$ possibilities, and so on.

Next, consider forests with k trees, with roots $A_1, \ldots, A_k$. By the same argument as above, the number of coalescent trees with root A_i is $n_i!(n_i-1)!/2^{n_i-1}$. The total number of forests is then the product of the number of these trees,

$$\prod_{j=1}^{k} \frac{n_j!(n_j-1)!}{2^{n_j-1}}. \tag{2.3}$$

Considering a single forest, in the context of combining its trees to form a single root $[n]$ coalescence tree, we must order the heights of each coalescence event. Within each tree, the order is determined, but between trees they are not. The total number of ways to order the coalescence events within this forest is therefore

$$\frac{\left(\sum_{j=1}^{k} n_j - 1\right)!}{\prod_{j=1}^{k}(n_j-1)!} = \frac{(n-k)!}{\prod_{j=1}^{k}(n_j-1)!}. \tag{2.4}$$

Additionally, in combining these trees, we must also determine the order which coalescence events occur above $A_1, \ldots, A_k$. This is simply a coalescence tree with k leaves, having

$$\frac{(k)!(k-1)!}{2^{k-1}} \tag{2.5}$$

possible orderings.

If we multiply the number of forests (2.3) by the number of intertree event orderings (2.4) and the number of ways to arrange the events above $A_1, \ldots, A_k$ (2.5), we obtain the number of root $[n]$ coalescence trees which contain the exact partition of $[n]$ into sets $A_1, \ldots, A_k$. Remembering that all trees have an equal probability under the Kingman coalescent, we divide by the overall number of root $[n]$ coalescence trees (2.2) to obtain the stated theorem. □

Corollary 2.2. *Suppose the set* $\{\{1\}, \ldots, \{n\}\}$ *undergoes* k *coalescences resulting in a partition of* $[n]$ *into* $n - k$ *sets. The probability 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* $k \geq 2$*, then additionally*

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

Proof. The probability $q(k, n)$ is zero if $2k > n$ because a partition of size 3 or more is inevitable after so many coalescences. As $(n-k)^{\underline{k}} = 0$ in this case, the formula holds.

Supposing $2k \leq n$, if a partition into $n - k$ sets has only sets of size 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 given by the number of ways to choose the $2k$ elements belonging to the sets of size 2 times the number of ways to pair those $2k$ elements,

$$\binom{n}{2k} \frac{(2k)!}{2^k k!}. \tag{2.6}$$

By Theorem 2.1, the probability of each of these partitions is equal to

$$\frac{k!(n-k)!(n-k-1)!}{n!(n-1)!} 2^k. \tag{2.7}$$

The formula for $q(k,n)$ is found by multiplying (2.6) and (2.7) and simplifying.

The bounds for $q(k,n)$ follow from calculations that are rather tedious.

First, if we let $f(\alpha) = \log(1-\alpha)$, we can use the Taylor expansion about 0 with remainder to obtain

$$\begin{aligned}\log(1-\alpha) &= -\alpha + \int_0^\alpha f^{(2)}(t)(\alpha - t)\,dt = -\alpha + \int_0^\alpha \frac{t-\alpha}{(1-t)^2}\,dt \\ &= -\alpha - \alpha^2 \int_0^1 \frac{1-s}{(1-\alpha s)^2}\,ds.\end{aligned}$$

If $\alpha \in [0, \frac{1}{2}]$,

$$\frac{1}{2} = \int_0^1 1 - s\,ds \leq \int_0^1 \frac{1-s}{(1-\alpha s)^2}\,ds \leq \int_0^1 \frac{1-s}{1-s/2}\,ds \leq 1.$$

Therefore,

$$\forall \alpha \in [0, \frac{1}{2}], \quad \exists u \in [\frac{1}{2}, 1] \quad \text{such that} \quad \log(1-\alpha) = -\alpha - u\alpha^2. \tag{2.8}$$

We will need to construct two more approximations to finish this proof. For the first, using a Riemann sum approximation for $\int_m^n \frac{1}{x}\,dx$, we get

$$\sum_{k=m}^{n-1} \frac{1}{k} = \log\left(\frac{n}{m}\right) + \text{ error}. \tag{2.9}$$

Considering the error term by term, the error for the first term would be given by $u_m(\frac{1}{m} - \frac{1}{m+1})$, for some $u_m \in [0,1]$. Summing the error of all terms, and considering it as a weighted sum of $u_i \in [0,1]$, we get

$$\frac{\text{error}}{\left(\frac{1}{m} - \frac{1}{n}\right)} = \frac{\sum_{k=m}^{n-1}\left(\frac{1}{k} - \frac{1}{k+1}\right)u_k}{\sum_{k=m}^{n-1}\left(\frac{1}{k} - \frac{1}{k+1}\right)} = u, \quad \text{where} \quad u \in [0,1].$$

Using this, we can then rewrite (2.9) as

$$\sum_{k=m}^{n-1} \frac{1}{k} = \log\left(\frac{n}{m}\right) + u\left(\frac{1}{m} - \frac{1}{n}\right), \quad \text{for some} \quad u \in [0,1]. \tag{2.10}$$

With an identical argument involving a Riemann sum approximation for $\int_m^n \frac{1}{x^2}\,dx$, we can find that

$$\sum_{k=m}^{n-1} \frac{1}{k^2} = \left(\frac{1}{m} - \frac{1}{n}\right) + u\left(\frac{1}{m^2} - \frac{1}{n^2}\right), \quad \text{for some} \quad u \in [0,1]. \tag{2.11}$$

Let $p = q(k, n)$. From the formula for $q(k, n)$ and 2.8, we have

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

for some $u \in [\frac{1}{2}, 1]$. The application of (2.8) is justified because $3k \leq n$ implies $k/(n-k+1) < 1/2$. Applying (2.10) and (2.11) to the two sums, we obtain

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

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

Thus,

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

The second inequality is obtained using (2.8). The third inequality notes $3k \leq n$ to conclude

$$\frac{1}{n-k+1} \leq \frac{1}{n-k} = \frac{3}{3n-3k} \leq \frac{3}{3n-n} = \frac{3}{2n}.$$

To prove the upper bound, we argue

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

The second inequality recognizes that $-x$ is a global overestimate of $\log(1-x)$, and the final inequality requires that $k \geq 2$. □

Lemma 2.3. *Consider the application of a single backward WF step to a sample of size n with a parental population of size N. Let $\mathcal{M}_n$ be the event that a merger of any kind has occurred. Then assuming $n \leq \sqrt{2N}$,*

$$\mathbb{P}(\mathcal{M}_n) \geq \frac{3}{10} \times \frac{n(n-1)}{N}.$$

Proof. A merger having occurred is the complement of the event that no merger has occurred, where each sample selects a unique parent. Therefore

$$\mathbb{P}(\mathcal{M}_n) = 1 - \prod_{k=1}^{n-1} \frac{N-k}{N} = 1 - \prod_{k=1}^{n-1} \left(1 - \frac{k}{N}\right).$$

For $k < N$, $1 - \frac{k}{N} \leq e^{-k/N}$, so

$$\mathbb{P}(\mathcal{M}_n) \geq 1 - \prod_{k=1}^{n-1} e^{-k/N} = 1 - e^{\frac{-n(n-1)}{2N}}$$

Additionally, so long as $x \in [0,1]$, $(1-e^{-1})x \leq 1 - e^{-x}$. Because $n \leq \sqrt{2N}$, $\frac{n(n-1)}{2N} < 1$. Therefore

$$\mathbb{P}(\mathcal{M}_n) \geq \left(1 - e^{-1}\right) \frac{n(n-1)}{2N} \geq \frac{3}{10} \times \frac{n(n-1)}{N}.$$

□

Lemma 2.4. *Consider the application of a single backward WF step to a sample of size n with parental population of size N. Let $\mathcal{D}_n^{(1)}$ be the event that more than one binary merger has occurred and/or a triple merger has occurred. Again, let $\mathcal{M}_n$ be the event that a merger of any kind has occurred. Then, as long as $n \leq \sqrt{2N}$,*

$$\mathbb{P}(\mathcal{D}_n^{(1)} | \mathcal{M}_n) \leq \frac{5}{3}\left(\frac{(n-2)(n-1)}{4N}\right) \leq \frac{n^2}{2N}.$$

Proof. $\mathcal{D}_n^{(1)}$ consists of the union of two events:

- The event where two or more binary mergers occur. Call this A.

- The event where a triple merger occurs. Call this B.

$\mathcal{D}_n^{(1)} = A \cup B \subset \mathcal{M}_n$, so

$$\mathbb{P}(\mathcal{D}_n^{(1)}|\mathcal{M}_n) = \frac{\mathbb{P}(A \cup B)}{\mathbb{P}(\mathcal{M}_n)} \leq \frac{\mathbb{P}(A) + \mathbb{P}(B)}{\mathbb{P}(\mathcal{M}_n)}. \tag{2.12}$$

We can overestimate $\mathbb{P}(A)$ by counting the ways to select two pairs of samples and multiplying this by the probability that given a selection of two pairs the samples within each pair would select the same parent, $1/N^2$. We obtain

$$\mathbb{P}(A) \leq \binom{n}{4} \times \frac{4!}{2^2 2!} \times \frac{1}{N^2} = \frac{n(n-1)(n-2)(n-3)}{8N^2}.$$

Similarly, we can overestimate $\mathbb{P}(B)$ by counting the ways to select a set of three samples and multiplying this by the probability that three samples would all select the same parent, $1/N^2$, obtaining

$$\mathbb{P}(B) \leq \binom{n}{3} \times \frac{1}{N^2} \leq \frac{n(n-1)(n-2)}{6N^2}.$$

Substituting these as well as our lower bound for $\mathbb{P}(\mathcal{M}_n)$ from Lemma 2.3 into 2.12, we find

$$\begin{aligned}
\mathbb{P}(\mathcal{D}_n^{(1)}|\mathcal{M}_n) &\leq \frac{\frac{n(n-1)(n-2)(n-3)}{8N^2} + \frac{n(n-1)(n-2)}{6N^2}}{\frac{3}{10}n(n-1)/N} \\
&= \frac{5}{3}\left(\frac{(n-2)(n-3)}{4N} + \frac{n-2}{3N}\right) \\
&\leq \frac{5}{3}\left(\frac{(n-2)(n-3)}{4N} + \frac{2(n-2)}{4N}\right) \\
&= \frac{5}{3}\left(\frac{(n-2)(n-1)}{4N}\right).
\end{aligned}$$

□

Theorem 2.5. *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 $\mathcal{D}^{(1)}$ be the event that a sample of size n undergoes more than a single binary merger and/or a triple merger during one or more backward WF steps in its genealogy. Let $\tilde{\mathcal{D}}_k^{(1)}$ be the event that in the genealogy while the ancestral sample size is k, a double binary merger or triple merger occurs. Furthermore, let A_k be the event that the sample size is ever k in the genealogy.

Given A_k, the final WF step for which the ancestral sample size is k is the only such WF step for which a merger will occur. As multiple mergers cannot occur in steps where no merger occurred,

$$\mathbb{P}(\tilde{\mathcal{D}}_k^{(1)}|A_k) = \mathbb{P}(\mathcal{D}_k^{(1)}|\mathcal{M}_k).$$

Additionally, $\tilde{\mathcal{D}}_k^{(1)} \subset A_k$, so we have

$$\mathbb{P}(\tilde{\mathcal{D}}_k^{(1)}) = \mathbb{P}(\tilde{\mathcal{D}}_k^{(1)}|A_k)\mathbb{P}(A_k) = \mathbb{P}(\mathcal{D}_k^{(1)}|\mathcal{M}_k)\mathbb{P}(A_k) \leq \mathbb{P}(\mathcal{D}_k^{(1)}|\mathcal{M}_k).$$

Clearly, $\mathcal{D}^{(1)} = \cup_{k=3}^{n} \tilde{\mathcal{D}}_k^{(1)}$. Therefore,

$$\mathbb{P}(\mathcal{D}^{(1)}) \leq \sum_{k=3}^{n} \mathbb{P}(\tilde{\mathcal{D}}_k^{(1)}) \leq \sum_{k=3}^{n} \mathbb{P}(\mathcal{D}_k^{(1)}|\mathcal{M}_k).$$

Substituting our upper bound from Lemma 2.4, we get

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

Assuming $n = N^{1/3-\epsilon}$,

$$\lim_{N\to\infty} \mathbb{P}(\mathcal{D}^{(1)}) \leq \lim_{N\to\infty} \frac{N^{-3\epsilon}}{2} = 0.$$

In the complement of $\mathcal{D}^{(1)}$, every merger is a single binary merger. □

Lemma 2.6. *Consider the application of a single backward WF step to a sample of size n with parental population of size N. Let $\mathcal{D}_n^{(c)}$ be the event that $c+1$ or more*

binary mergers have occurred and/or a triple merger has occurred. Again, let $\mathcal{M}_n$ be the event that a merger of any kind has occurred. Then, as long as $n \leq \sqrt{2N}$,

$$\mathbb{P}(\mathcal{D}_n^{(c)}|\mathcal{M}_n) \leq \frac{5}{3}\left(\frac{n^{2c}}{2^c(c+1)!N^c} + \frac{n}{3N}\right) \leq \frac{n^{2c}}{2^{c-1}(c+1)!N^c} + \frac{n}{N}.$$

Proof. $\mathcal{D}_n^{(c)}$ consists of the union of two events:

- The event where $c+1$ or more binary mergers occur. Call this A.
- The event where a triple merger occurs. Call this B.

$\mathcal{D}_n^{(c)} = A \cup B \subset \mathcal{M}_n$, so

$$\mathbb{P}(\mathcal{D}_n^{(c)}|\mathcal{M}_n) = \frac{\mathbb{P}(A \cup B)}{\mathbb{P}(\mathcal{M}_n)} \leq \frac{\mathbb{P}(A) + \mathbb{P}(B)}{\mathbb{P}(\mathcal{M}_n)}. \tag{2.13}$$

We can overestimate $\mathbb{P}(A)$ by counting the ways to select $c+1$ pairs of samples and multiplying this by the probability that given a selection of $c+1$ pairs the samples within each pair would select the same parent, $1/N^{c+1}$. We obtain

$$\mathbb{P}(A) \leq \binom{n}{2c+2} \times \frac{(2c+2)!}{2^{c+1}(c+1)!} \times \frac{1}{N^{c+1}} \leq \frac{n^{2c+1}(n-1)}{2^{c+1}(c+1)!N^{c+1}}.$$

Similarly, we can overestimate $\mathbb{P}(B)$ by counting the ways to select a set of three samples and multiplying this by the probability that three samples would all select the same parent, $1/N^2$, obtaining

$$\mathbb{P}(B) \leq \binom{n}{3} \times \frac{1}{N^2} \leq \frac{n^2(n-1)}{6N^2}.$$

Substituting these as well as our lower bound for $\mathbb{P}(\mathcal{M}_n)$ from Lemma 2.3 into 2.13, we find

$$\begin{aligned}\mathbb{P}(\mathcal{D}_n^{(1)}|\mathcal{M}_n) &\leq \left(\frac{n^{2c+1}(n-1)}{2^{c+1}(c+1)!N^{c+1}} + \frac{n^2(n-1)}{6N^2}\right) \times \frac{10}{3} \times \frac{N}{n(n-1)} \\ &= \frac{5}{3}\left(\frac{n^{2c}}{2^c(c+1)!N^c} + \frac{n}{3N}\right).\end{aligned}$$

□

Theorem 2.7. *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 $\mathcal{D}^{(c)}$ be the event that a sample of size n undergoes more than $c+1$ binary mergers and/or a triple merger during one or more backward WF steps in its genealogy. Let $\tilde{\mathcal{D}}_k^{(c)}$ be the event that in the genealogy while the ancestral sample size is k, $c+1$ binary mergers or a triple merger occurs. Furthermore, let A_k be the event that the sample size is ever k in the genealogy.

Given A_k, the final WF step for which the ancestral sample size is k is the only such WF step for which a merger will occur. As multiple mergers cannot occur in steps where no merger occurred,

$$\mathbb{P}(\tilde{\mathcal{D}}_k^{(c)}|A_k) = \mathbb{P}(\mathcal{D}_k^{(c)}|\mathcal{M}_k).$$

Additionally, $\tilde{\mathcal{D}}_k^{(c)} \subset A_k$, so we have

$$\mathbb{P}(\tilde{\mathcal{D}}_k^{(c)}) = \mathbb{P}(\tilde{\mathcal{D}}_k^{(c)}|A_k)\mathbb{P}(A_k) = \mathbb{P}(\mathcal{D}_k^{(c)}|\mathcal{M}_k)\mathbb{P}(A_k) \leq \mathbb{P}(\mathcal{D}_k^{(c)}|\mathcal{M}_k).$$

Clearly, $\mathcal{D}^{(c)} = \cup_{k=3}^{n} \tilde{\mathcal{D}}_k^{(c)}$. Therefore,

$$\mathbb{P}(\mathcal{D}^{(c)}) \leq \sum_{k=3}^{n} \mathbb{P}(\tilde{\mathcal{D}}_k^{(c)}) \leq \sum_{k=3}^{n} \mathbb{P}(\mathcal{D}_k^{(c)}|\mathcal{M}_k).$$

Substituting our upper bound from Lemma 2.6, assuming $n \leq \sqrt{2N}$, we get

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

Assuming $n = N^{\frac{c}{2c+1}-\epsilon}$,

$$\lim_{N\to\infty} \mathbb{P}(\mathcal{D}^{(c)}) \leq \lim_{N\to\infty} \frac{N^{-c\epsilon}}{2^{c-1}(c+1)!} + N^{-1/(2c+1)-2\epsilon} = 0.$$

□

We now turn to the sample frequency spectrum under Wright Fisher. Unlike the approach in [29, 5], our approach does not look at the internal structure of the genealogical tree.

Let $\mathcal{A}_n$ denote the condition that the genealogy of a sample of size n involves exactly one mutation under 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 $q(n, N)$ 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 multiple 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 triplet 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

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

Note that we are making the usual assumption that the sample cannot be hit with both a mutation and a merger in the same generation. This assumption is rational for the following reasons: First, we limit ourselves to samples of size $N^{1/3-\epsilon}$ or less. Second, the condition $\mathcal{A}_n$ 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 $\mathcal{A}_n \cap \mathcal{B}_n$, 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 $\mathcal{A}_n \cap \mathcal{B}_n$ in the limit of zero mutation.

Let $f(j,n)$ be the probability that j out of n samples are mutants under the condition $\mathcal{A}_n \cap \mathcal{B}_n$. This probability can be calculated exactly using the following recurrence[3] :

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

In this recurrence, we have used Knuth's notation [24, 53] by which $[j=1]$ evaluates to 1 if $j=1$ and 0 otherwise.

[3]To see where this recurrence comes from, let us examine the case of calculating $f(2,6)$ given a knowledge of $f(j,5)$. We have a sample of size 6. There are two possibilities. Either the most recent event is a coalescence or a mutation. If it is a mutation, then the number of mutants is necessarily 1, so we ignore this possibility. If it is a coalescence (an event with probability $(1-\mu_6)$), then when the ancestral sample size was 5, the genealogy contained exactly one mutation and so we know the probability of each potential number of mutants. The two prior states which could lead to $(2,6)$ are $(1,5)$ where the mutant splits into two lines (probability $1/5$ to select the mutant and $f(1,5)$ that it was in that state) or $(2,5)$ where a non mutant splits into two lines (probability $3/5$ to select a non mutant and $f(2,5)$ that it was in that state). Therefore the probability of two mutants in a sample of size 6 is

$$f(2,6) = (1-\mu_6)\left(f(1,5)\frac{1}{5} + f(2,5)\frac{3}{5}\right).$$

To obtain the classic 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 only one mutation in the genealogy. The exact solution of the recurrence

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

is given by

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

Lemma 2.8. *Consider the application of a single backward WF step to a sample of size n with parental population of size N. Letting $\tilde{A}_{12}$ denote the event that samples 1 and 2 merge and there is no other merger,*

$$\mathbb{P}(\tilde{A}_{12}) \geq \frac{1}{N}\left(1-\frac{n^2}{2N}\right)$$

Proof. Let A_{12} be the event that samples 1 and 2 merge under the backward WF step: $\mathbb{P}(A_{12}) = \frac{1}{N}$.

Let $A_{12}^{(t)}$ be the event that 1 and 2 merge and one of the other $(n-2)$ samples has the same parent of 1 and 2. Then $A_{12}^{(t)} = \cup_{j=3}^{n} A_{12j}$ where A_{12j} is the event where 1, 2 and j have the same parent. Because $\mathbb{P}(A_{12j}) = \frac{1}{N^2}$,

$$\mathbb{P}\left(A_{12}^{(t)}\right) = \mathbb{P}\left(\cup_{j=3}^{n} A_{12j}\right) \leq \frac{(n-2)}{N^2}.$$

Let $A_{12}^{(d)}$ be the event that 1 and 2 merge, and that there is some other pair that merges. We have $A_{12}^{(d)} = \cup_{j,k} A_{12,jk}$, where $A_{12,jk}$ is the event that 1 and 2 as well as j and k have the same parent. The union is over all combinations of j and k that are not 1 or 2. Because $\mathbb{P}(A_{12,jk}) \leq \frac{1}{N^2}$,

$$\mathbb{P}\left(A_{12}^{(d)}\right) = \mathbb{P}\left(\cup_{j,k} A_{12,jk}\right) \leq \binom{(n-2)}{2}\frac{1}{N^2}.$$

As $\tilde{A}_{12} = A_{12} - A_{12}^{(t)} - A_{12}^{(d)}$,

$$\mathbb{P}(\tilde{A}_{12}) \geq \frac{1}{N} - \frac{(n-2)}{N^2} - \frac{(n-2)(n-3)}{2N^2} \geq \frac{1}{N}\left(1 - \frac{n^2}{2N}\right).$$

□

Lemma 2.9. *For $n < \sqrt{2N}$,*

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

Proof. We will use our notation from Theorem 2.5, where $\mathcal{D}_n^{(1)}$ represents a multiple binary merger or a triple merger in a single backward WF step, and $\mathcal{M}_n$ represents a merger of any kind in a single backward WF step. For ease of notation, let $E = (\mathcal{D}_n^{(1)})^c$.

It is easy to bound above the probability of $\mathcal{M}_n$ and $\mathcal{D}_n^{(1)}$. For $\mathcal{M}_n$, count the number of ways we can select two merging samples and multiply by the probability that two specific samples would have the same parent. We get

$$\mathbb{P}(\mathcal{M}_n) \leq \binom{n}{2}\frac{1}{N} = \frac{n(n-1)}{2N}. \tag{2.14}$$

Using the same technique, as used in the proof of Theorem 2.5, we get

$$\mathbb{P}(\mathcal{D}_n^{(1)}) \leq \binom{n}{4} \times 3 \times \frac{1}{N^2} + \binom{n}{3} \times \frac{1}{N^2} \leq \frac{n^4}{8N^2}. \tag{2.15}$$

Finally, use $\tilde{A}_{ij}$ as in Lemma 2.8 to denote the event that only samples i and j merge in a single WF step. The event that some two merge and there is no other merger, $\mathcal{M}_n \cap E$ is equivalent to the union of $\tilde{A}_{ij}$ over all possible i and j. Notice that $\{\tilde{A}_{ij}\}$ are disjoint events. Using the bound from Lemma 2.8, we get

$$\mathbb{P}(\mathcal{M}_n \cap E) = \mathbb{P}\left(\cup_{i,j}\tilde{A}_{ij}\right) = \binom{n}{2}\mathbb{P}\left(\tilde{A}_{12}\right) \geq \frac{n(n-1)}{2N}\left(1 - \frac{n^2}{2N}\right). \tag{2.16}$$

We see that

$$q(n,N) = \mathbb{P}(\mathcal{M}_n|E) = \frac{\mathbb{P}(\mathcal{M}_n \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|\mathcal{M}_n)\mathbb{P}(\mathcal{M}_n)}{\mathbb{P}(E)}.$$

To obtain the upper bound, notice that $\mathbb{P}(E) = 1 - \mathbb{P}(\mathcal{D}_n^{(1)})$ and substitute the upper bounds from 2.14 and 2.15, getting

$$\begin{aligned} q(n,N) = \frac{\mathbb{P}(E|\mathcal{M}_n)\mathbb{P}(\mathcal{M}_n)}{\mathbb{P}(E)} &\leq \frac{\mathbb{P}(\mathcal{M}_n)}{1-\mathbb{P}(\mathcal{D}_n^{(1)})} \leq \frac{n(n-1)/2N}{1-n^4/(8N^2)} \\ &\leq \frac{n(n-1)}{2N}\left(1+\frac{n^4}{4N^2}\right). \end{aligned}$$

The final inequality requires that $n^4/(8N^2) \leq 1/2$ which follows from our assumption that $n \leq \sqrt{2N}$.

To see the lower bound simply take our bound from 2.16 to get

$$q(n,N) = \frac{\mathbb{P}(\mathcal{M}_n \cap E)}{\mathbb{P}(E)} \geq \mathbb{P}(\mathcal{M}_n \cap E) \geq \frac{n(n-1)}{2N}\left(1-\frac{n^2}{2N}\right).$$

□

Lemma 2.10. *For $n < \sqrt{N}$,*

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

Proof. Take $q(j,N) = \frac{j(j-1)}{2N}(1-s_j)$. Then by the previous lemma,

$$s_j \in [-j^4/4N^2, j^2/2N].$$

Using the definition of μ_n, we get

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

To obtain the lower bound, use $s_j \geq -j^4/4N^2 \geq -n^4/4N^2$ in the denominator and $s_n \leq n^2/2N$ in the numerator, resulting in

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

To obtain the upper bound, use $s_j \le j^2/2N \le n^2/2N$ in the denominator and $s_n \ge -n^4/4N^2$ in the numerator. We get

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

The final inequality requires $n^2 \le N$. □

Lemma 2.11. *If* $n \le N^{1/3-\epsilon}$,

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

Proof. Recall

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

and that $\tilde{f}(j,n)$ is obtained using the same recurrence, but replacing μ_n with $\tilde{\mu}_n$.

Note that $|ab - \tilde{a}\tilde{b}| \le |a-\tilde{a}||b| + |b-\tilde{b}||\tilde{a}|$. With this, for $j = 2, \ldots, n-1$ we see

(2.17)

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

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

Next we will sum the above over all j. Notice that

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

Therefore, when summing 2.17 over all j, we obtain

$$\sum_{j=1}^{n-1}\left|f(j,n)-\tilde{f}(j,n)\right| \leq 3|\mu_n-\tilde{\mu}_n|+(1-\tilde{\mu}_n)\sum_{j=1}^{n-2}\left|f(j,n-1)-\tilde{f}(j,n-1)\right|. \tag{2.18}$$

The 2 is changed to 3 to allow for the extra $|\mu_n-\tilde{\mu}_n|$ in the $j=1$ case. For ease of notation, let

$$S(n-1)=\sum_{j=1}^{n-1}\left|f(j,n)-\tilde{f}(j,n)\right|.$$

Using the fact that $(1-\tilde{\mu}_n)\leq 1$, and that $f(1,2)=\tilde{f}(1,2)=1$, from 2.18 we see that

$$\begin{aligned}
S(n-1) &\leq 3|\mu_n-\tilde{\mu}_n|+(1-\tilde{\mu}_n)S(n-2)\\
&\leq 3|\mu_n-\tilde{\mu}_n|+S(n-2)\\
&\leq 3|\mu_n-\tilde{\mu}_n|+3|\mu_{n-1}-\tilde{\mu}_{n-1}|+S(n-3)\\
&\leq \sum_{k=3}^{n}|\mu_k-\tilde{\mu}_k|+S(3)\\
&= \sum_{k=3}^{n}|\mu_k-\tilde{\mu}_k|.
\end{aligned}$$

Additionally, by Lemma 2.10, for $n\leq\sqrt{N}$,

$$|\mu_k-\tilde{\mu}_k|\leq\frac{n^2}{N}+\frac{n^4}{4N^2}+\frac{n^6}{4N^3}.$$

Therefore,

$$\begin{aligned}
\sum_{j=1}^{n-1}\left|f(j,n)-\tilde{f}(j,n)\right| &\leq \sum_{k=3}^{n}\frac{k^2}{N}+\frac{k^4}{4N^2}+\frac{k^6}{4N^3}\\
&\leq \frac{n^3}{N}+\frac{n^5}{4N^2}+\frac{n^7}{4N^3}.
\end{aligned}$$

Using our assumption that $n\leq N^{1/3-\epsilon}$, we see that

$$\lim_{N\to\infty}\frac{1}{2}\sum_{j=1}^{n-1}\left|f(j,n)-\tilde{f}(j,n)\right| \leq \lim_{N\to\infty}\left(N^{-3\epsilon}+\frac{1}{4}N^{-1/3-5\epsilon}+\frac{1}{4}N^{-3/3-7\epsilon}\right)=0.$$

□

Theorem 2.12. *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 2.5, the probability that any backward WF step produces a simultaneous binary merger or a triple merger converges to zero as $N\to\infty$. Thus in the limit of large N we may assume the condition $\mathcal{B}_n$. Under this condition,

$$f(j,n)=f_{WF}(j,n)\quad\text{and}\quad\tilde{f}(j,n)=\frac{1/j}{\mathcal{H}_{n-1}}.$$

Therefore, because we are assuming $n\leq N^{1/3-\epsilon}$, we may invoke Lemma 2.11 to infer this theorem. □

2.6 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. In fact, the probability of such events in a single backward WF step is greater than zero. However, by Theorem 2.5 the probability 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 will derive an algorithm which calculates the probability that the WF genealogy involves only single binary mergers. Additionally we modify that algorithm to calculate the probability that the WF genealogy does not include 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 WF step. Then

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

where N is 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 N. Then

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

for $0 \leq 2k \leq n$. The formulas are valid even for $n \geq N$. This formula can be justified as follows. First, we choose $2k$ samples which will participate in k simultaneous binary mergers. These may be selected in $\binom{n}{2k}$ ways. To group the $2k$ samples into k pairs, the first sample may be paired in $(2k-1)$ ways, the second of the remaining $(2k-2)$ samples may be paired in $(2k-3)$ ways and so on. 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.

2.6.1 Probability of at most a single binary merger in any generation

For the current generation from which a sample of size 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 $\phi_n(k, t)$ 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 $k = 1, \ldots, N(t)$. When $k = 0$, $\phi_n(k, t)$ has a special interpretation: $\phi_n(0, t)$

is the probability that the WF genealogy from the current generation to ancestral generation t contains 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 $k \neq n$.

Suppose the data at time t is $\phi_n(k, t)$. The crux of the algorithm is to generate data at time $t + 1$, and the recurrence

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

does that for $k = 1, \ldots, \min(n, N(t+1))$. If the size of the multiple merger free 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 a larger change then that would necessitate a multiple merger. The two possibilities are disjoint, and the recurrence sums over those two possibilities. The recurrence for $\phi_n(k, t)$ is similar in structure to equation (3) in the appendix of [5]. The recurrences for genealogical quantities (as well as for the sample frequency spectrum viewed from a genealogical perspective) generally have a similar form [84].

The quantity $\phi_n(0, t + 1)$, with its special interpretation, is calculated using

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

where $n^* = \min(n, N(t + 1))$. Notice that this absorbs all of the cases where $k > n^* + 1$.

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 that some backward WF step involves a multiple merger is greater than 0.9999.

The probability of a multiple merger between ancestral generations t and $t + 1$ conditioned on at at most single binary mergers in backward WF steps preceding t

is given by

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

This formula is used to visualize the effect of bottlenecks.

2.6.2 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 $\psi_n(k,t)$ 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 that a triple merger occurs 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 $k\neq n$.

Suppose the data at time t is $\psi_n(k,t)$. The recurrence

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

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, denoted above 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,

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

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 to

$t+1$ conditioned on no triple merger from 0 to t is

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

Again, we will use this to visualize the effect of bottlenecks.

As t increases, a probability such as $\psi_n(n,t)$ becomes quite small but remains positive. Holding on to theses tiny numbers makes the algorithm quite expensive for large sample sizes. This algorithm (as well as 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. If probabilities smaller than ϵ_{tol} are ignored, there is a rapid reduction in the ancestral sample sizes that are tracked at ancestral generation t in the algorithm if n is large. The total contribution of $\psi_n(k,t)$ to probabilities in all later stages is bounded by $\psi_n(k,t)$ because the recurrence sums over disjoint possibilities. As a result, the total error caused by ignoring probabilities below ϵ_{tol} is 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 the ignored probability is vanishingly small even with $n = G = 10^{20}$.

CHAPTER 3

The Wright-Fisher Site Frequency Spectrum as a Perturbation of the Coalescent's[1]

3.1 Introduction

An attractive aspect of genealogical analysis is that it begins with current samples whose sequence data are directly measured. Wright-Fisher (WF) and the coalescent are two theoretical models used to make deductions about the genealogies of the current samples [11].

The coalescent was derived and justified by [52] as an approximation of the WF model. Kingman's analysis and extensions by other authors [62, 63] assume the current sample size n to be fixed as the haploid population size N becomes large.

In view of the rapid increase in sample sizes in human genetics (see [44], for example), it is worth asking how close the WF and coalescent models are for large samples. A key property of the coalescent is that the genealogy is constructed entirely using binary mergers. In Chapter 2, we have shown that for sample sizes $n = o\left(N^{1/3}\right)$, WF genealogies involve only binary mergers with probability tending to 1. A more precise result derived here states that if the sample size is given by $n = \alpha N^{1/3}$, the probability that the WF genealogy involves only binary mergers is $\exp\left(-\frac{\alpha^3}{12}\right)$ in the large N limit. To understand the onset of the deviation of WF from the coalescent, [20] as well as [5] looked at triple mergers, where three individuals merge

[1] A modified version of this chapter, under the same title, is published in *Theoretical Population Biology* [60]

into a common parent over a single WF generation. Among other results, we show that if the sample size is $n = \alpha N^{1/2}$, the expected number of triple mergers in the WF genealogy is $\alpha^2/6 + (\exp(-\alpha^2/2) - 1)/3$. This last result is in agreement with the $N^{1/2}$ scaling deduced in Chapter 2.

With regard to sequence data, such results are perhaps too exacting. Detailed agreement in the genealogy is essential to reproduce the Kingman partition distribution [51] at each step of the genealogy. However, summary statistics such as the site frequency spectrum are not so refined. The site frequency spectrum of a sample of size n, which may be directly obtained from sequence data, consists of the probability that j of the samples are mutants and $n - j$ are ancestral, $j = 1, \ldots, n-1$, at a base pair. The site is assumed to be polymorphic with a single mutation at some individual that is an ancestor of some but not all samples.

The site frequency spectrum has been widely used for making demographic inferences (see [15, 19, 29, 30, 43, 45, 55, 90], for example) and is therefore a good basis to understand the difference between WF and the coalescent with regard to sequence data. If the genealogy is given by the coalescent and if μ is the mutation rate per site per generation, the probability that j out of n samples are mutants is

$$\frac{1/j}{\mathcal{H}_{n-1}}, \tag{3.1}$$

where $\mathcal{H}_{n-1} = 1 + \frac{1}{2} + \cdots + \frac{1}{n-1}$ is the harmonic number[2], assuming μ to be so small that μN is negligible and assuming the sample to be polymorphic at the site. We will derive the first perturbing term that follows (3.1) under the assumption of WF genealogy.

The elegant formula (3.1) for the probability of j mutants has a long and complicated history. [17] stated that the correlation between heights of fathers and sons

[2]In sources such as [21], the harmonic number $\mathcal{H}_{n-1}$ is denoted by a_n. The notation we use is from [24].

was 0.5 and attempted to obtain a Mendelian explanation of that correlation. He was thus led to a consideration of "gene ratios," which is equivalent to counting the number of mutants. He derived the numerator of (3.1) some years later [18]. [95, p. 120] had contacted Fisher earlier, noting (among other discrepancies) that he obtained $2N \log(1.8N)$ for the size of the genealogy, whereas [17] had obtained $\sqrt{\pi} N^{3/2}$.[3] There can be little doubt that Fisher was aided by [95] in coming up with the arguments that led him to the numerator of (3.1) as well as another result we will review shortly.

The size of the genealogy under WF is equal to the number of ancestors (with the current sample included) with $1, \ldots, n-1$ (but not n) descendants in the sample; in other words, the number of ancestors who would make the sample polymorphic if hit with a mutation. The b-branch length of the genealogy is the number of ancestors with exactly b descendants for $b = 1, \ldots, n-1$. The size of the genealogy and the b-branch length are defined analogously for the coalescent, with the difference that the number of generations a lineage survives is no longer an integer. Ancestors of the current sample will be referred to as ancestral samples. Ancestors of the current sample in the same generation will be referred to as an ancestral sample. An ancestral sample induces a partition of the current sample, and for the coalescent, the partition follows the Kingman partition distribution as shown in Chapter 2 and in [11, 51, 29].

[48] (also see [49, p. 222]) solved the diffusion equation for gene frequencies. From that point, (3.1) can be derived, although an argument connecting mutant frequencies in the population to that in the sample (such as the argument in [11, p. 51]) would be needed. The first such argument was given by [14], who also introduced the "frequency spectrum" terminology. A coalescent derivation of (3.1) was given

[3] Wright's result is the same as the modern coalescent estimate of the expected size of the genealogy, with 1.8 being his approximation to e^{γ}, where γ is Euler's constant. Wright's $2N$ is the same as our N and his μ is the same as our 2μ. We have modified his formulas accordingly.

by [21], as a consequence of the expectations and variances of b-branch lengths of the coalescent genealogy and was preceded by the treatment of special cases [22, 83]. A mathematically complete treatment, allowing for varying population sizes, is due to [29], [73] and [74]. A concise and elegant approach to the main ideas of [73] was obtained recently by [93].

If the genealogy is given by WF, we show that the probability of j mutants out of n is given by

$$\begin{aligned} &\frac{1/j}{\mathcal{H}_{n-1}} - \frac{1}{6N\mathcal{H}_{n-1}(n-1)} - \frac{(j-1)}{6N\mathcal{H}_{n-1}(n-1)(n-j)} \\ &\qquad + \frac{1}{6N\mathcal{H}_{n-1}j} - \frac{n}{12N\mathcal{H}_{n-1}^2 j} + \frac{n[j=1]}{12N\mathcal{H}_{n-1}} + \cdots , \end{aligned} \tag{3.2}$$

where $[j = 1]$ is 1 if the assertion $j = 1$ is true and 0 otherwise and where $j = 1, \ldots, n-1$. The result (3.2) is perturbative in that it gives the N^{-1} terms but not the N^{-2} terms.

The main point in calculating the first terms of the perturbation series, shown in (3.2), is to understand the onset of deviations. Under WF, children are split between parents according to the multinomial distribution. Under the coalescent, the split is uniform (as shown in Chapter 2). The uniform split is intuitively unreasonable and appears implausible. For example, if two parent have ten children the splits $9+1$ and $5+5$ are equally likely. The assumption of at most a single binary merger per generation breaks down for sample sizes that are as small as $N^{1/3}$. Yet the first terms of the perturbative series (3.2) show that the deviation in the site frequency spectrum sets in only for sample sizes n that are order of the population size N.

The first few terms in the perturbative series cannot be a good approximation to the total deviation except for small n (however, see Figures 3.2 and 3.3). It is well-known that the first neglected term in a pwer series often gives a good idea of the error. In the same way, the $\frac{1}{N}$ terms in (3.2) give an idea of the various phenomena

at work in making the WF frequency spectrum differ from that of the coalescent. As noted by earlier authors [5, 20, 92], the WF frequency spectrum elevates the probability of singletons ($j = 1$ mutants) and lowers the probability of j mutants for each $j > 1$. Such a movement in mutant probabilities may be verified explicitly from the last two terms of (3.2), which are the only terms that increase with n. The last two terms increase approximately linearly with sample size. For population sized samples, (3.2) yields an estimate of $1/12\mathcal{H}_{N-1}$ (or $1/12 \log N$) for the amount by which singleton probability is raised under WF. Even with later terms in the perturbative series not taken into account, that estimate is off only by a factor of 3/2.

In principle, better approximations can be obtained by calculating more terms of the perturbative series. However, the extension of our method to calculate even the N^{-2} terms, which are presumably of the form n^2/N^2, appears difficult. Therefore, we give a separate analysis of population sized samples with $n = N$, with the work of [92] being our starting point. [18] gave an ingenious derivation of b-branch lengths of WF genealogies with $n = N$, although some of his arguments are not entirely clear.[4] [92] gave a different and more transparent argument for the $b = 1$ case, which we extend to $b > 1$.

If p_j and q_j are two probability distributions over $j = 1, \ldots, n-1$, the total variation (TV) distance between them is $\frac{1}{2}\sum_{j=1}^{n-1} |p_j - q_j|$. The total variation distance is the maximum difference in probabilities of any possible event under the two distributions [6, p. 126] and is therefore a quite robust way to compare probability distributions. The total variation distance between the frequency spectrums under

[4]Specifically, in deriving the functional equation $\phi(e^{x-1}) - \phi(x) = 1 - x$, [18, p. 209] assumes silently that most mutations do not become fixed in the population after assuming the probability of fixation to be 1/2 one paragraph back. That most mutations do not become fixed was known to [95], and Kimura later proved the probability of a neutral mutation becoming fixed to be $1/N$.

WF and the coalescent for a population sized sample with $n = N$ is approximately

$$\frac{0.1204}{\mathcal{H}_{N-1}} - \frac{0.1124}{\mathcal{H}_{N-1}^2} + \cdots,$$

with a slight change in the approximation for $N > 6.8 \times 10^5$. For $N = 2 \times 10^4$, the baseline assumption in human genetics [11], the total variation distance is only around 1%.

[20] has connected the greater speed of mergers under the coalescent to the elevation of singleton probability under WF. As we will explain, the coalescent is indeed faster for $n \ll N^{1/2}$ but for $n \approx N$, the picture is not so clear. We refer to the same phenomenon as a mismatch in rates of merger to cover both cases. Another difference between the models is in the way children are partitioned between the parents as discussed above. In particular, the offspring distribution is approximately Poisson for WF but approximately geometric for the coalescent.

To disentangle the two effects, we define an intermediate model called the discrete coalescent. In the discrete coalescent, the number of parents of a sample of size n has exactly the same distribution as in WF. However, once the number of parents is determined, the children are split between the parents according to Kingman's partition distribution [51]. The intermediate model shows that the effect of the mismatch in rates is twice as great as the effect of the difference in the way children are split between parents. The two effects are of opposite sense and combine to cause a reduction in overall error.

3.2 Poisson approximations to Wright-Fisher genealogies

In a backward WF step, each haploid individual chooses one out of N parents with equal probability and independently of all other individuals in its generation. The WF genealogy of a sample is built up using backward WF steps. The coalescent

[52] may be thought of as a rate varying Poisson approximation of WF genealogies.

Other Poisson approximations may be used to capture more detailed information about WF genealogies. The clumping heuristic is a general method for deriving Poisson approximations [4]. Applications of the heuristic require greater sophistication when the "clumps" are disconnected. In the case of WF, the clumps have a relatively simple form and the heuristic is not difficult to apply.

For the most part, the following basic fact is all that we will need. Suppose the probability of occurrence of an event (such as a thunderstorm) in the interval $(u, u + du)$ is $\lambda(u)\,du$. Then the total number of occurrences of the event in the domain $[a, b]$ has Poisson distributed with rate $\Lambda = \int_a^b \lambda(u)\,du$. In particular, the probability of k occurrences is $\frac{\Lambda^k}{k!}e^{-\Lambda}$. If an event is rare in every neighborhood, the total number of occurrences is approximately Poisson with the rate obtained by summing over the domain.

Let n be the number of samples and N the size of the parental generation. If δ is the number of samples lost due to mergers in a single backward WF step, the number of parental samples is $n - \delta$ and we have

$$\mathbb{E}\,\delta = n - N + N\left(1 - \frac{1}{N}\right)^n$$

$$\operatorname{Var}\delta = N\left(\left(1 - \frac{1}{N}\right)^n - \left(1 - \frac{2}{N}\right)^n\right) + N^2\left(\left(1 - \frac{2}{N}\right)^n - \left(1 - \frac{1}{N}\right)^{2n}\right) \tag{3.3}$$

[94]. When n is fixed, $\mathbb{E}\,\delta = \frac{n(n-1)}{2N} - \binom{n}{3}\frac{1}{N^2} + \cdots$ and $\operatorname{Var}\delta - \mathbb{E}\,\delta = -2n^3/3N^2 + \cdots$, suggesting a Poisson approximation for small n which turns out to be the Kingman coalescent. More generally, if $n = N^{\mathfrak{a}}$, where $\mathfrak{a} \in [0, 1)$, we have $\mathbb{E}\,\delta - \operatorname{Var}\delta = \mathcal{O}(N^{3\mathfrak{a}-2}) = o\,(N^{\mathfrak{a}})$, suggesting a Poisson approximation to δ for $\mathfrak{a} < 1$.

For $i = 1, \ldots, k$, the ith sample has the same parent as the $k + 1$st sample with probability $1/N$, which is a rare event for $N \gg 1$. The accumulated rate is k/N.

Thus, the probability the $k+1$st sample has the same parent as one of the prior samples is approximately $1-\exp(-k/N)$.

Suppose the number of samples is n. The probability that the $i+1$st sample merges with one of the prior samples is $1-\exp(-i/N)$ approximately, which is a rare event for $i \ll N$. For $n \ll N$, the cumulative rate $\sum_{i=1}^{n-1}(1-\exp(-i/N))$ is the left hand Riemann sum of the integral

$$\begin{aligned}\int_1^n \left(1-e^{-\frac{u}{N}}\right) du &= n-1+N\left(e^{-\frac{n}{N}}-e^{-\frac{1}{N}}\right)\\ &= n+N(e^{-\frac{n}{N}}-1)+\cdots\end{aligned}$$

We may correct for an error that occurs in replacing the sum by an integral by taking $\lambda_\delta(n) = n+N(\exp(-n/N)-1)-n/2N$. (The sum is now approximated to order N^{-2} .)

For $n \ll N$, δ approximately follows a Poisson distribution of rate $\lambda_\delta(n)$. Thus, $\mathbb{P}(\delta = k) \approx \exp(-\lambda_\delta(n)) \times \lambda_\delta(n)^k/k!$. In fact, $\mathbb{E}\delta = \lambda_\delta(n)+\epsilon$, where $\epsilon = n^2/2N^2+\cdots$ is of the same order as the error in the Poisson approximation.

3.2.1 Non-binary mergers

If the sample size is small enough, mergers in any generation are likely to be single binary mergers as in the Kingman coalescent. As the sample size increase, multiple binary mergers may appear with some likelihood and then triple mergers and so on (as seen in Chapter 2).

The probability of something other than a binary merger conditional on $\delta \geq 1$ is

$$\frac{1-e^{-\lambda_\delta(n)}-\lambda_\delta(n)e^{-\lambda_\delta(n)}}{1-e^{-\lambda_\delta(n)}}.$$

For $n \ll N^{1/2}$, $\lambda_\delta(n) = n^2/2N$ is a good approximation. Because non-binary mergers at onset are double binary mergers (as shown in Chapter 2), we have the cumulative

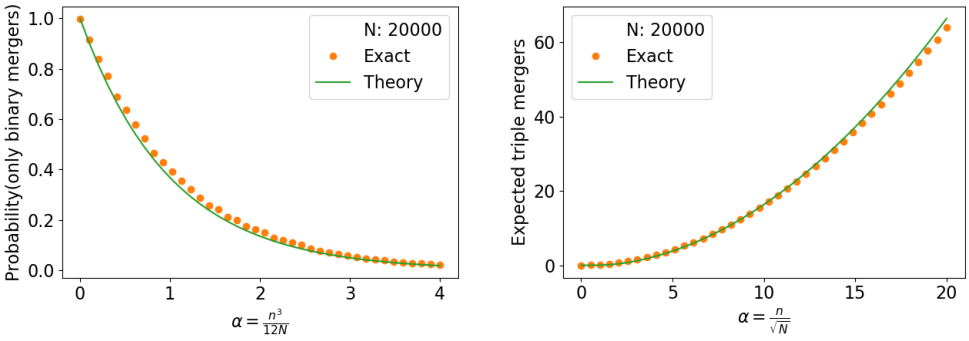


Figure 3.1: Plots verifying the approximations implied by (3.4) and (3.5). The exact numbers are from computer programs described in [5] and Chapter 2.

rate of double binary mergers (or non-binary mergers) to be

$$\begin{aligned}\Lambda_{2^2}(n) &= \int_0^n \frac{1 - e^{-x^2/2N} - (x^2/2N)e^{-x^2/2N}}{1 - e^{-x^2/2N}}\, dx \\ &= \int_0^n \frac{e^{x^2/2N} - 1 - x^2/2N}{e^{x^2/2N} - 1}\, dx \\ &= \frac{n^3}{12N} + \cdots, \end{aligned} \tag{3.4}$$

where the last step is from a power series expansion of $e^{x^2/2N}$. If $n = \alpha N^{1/3}$, $\Lambda_{2^2}(n) = \alpha^3/12$ implying the probability of coalescence with only binary mergers to be $1 - \exp(-\alpha^3/12)$ (see Figure 3.1) and the probability of exactly k binary mergers in the genealogy to be $\exp(-\beta)\beta^k/k!$, where $\beta = \alpha^3/12$.

3.2.2 Simultaneous binary mergers

The rate $\Lambda_{2^p}(n)$ for p simultaneous binary mergers is obtained similarly. A p-fold simultaneous binary merger occurs during a single backward WF step conditional on $\delta \geq 1$ with probability

$$\begin{aligned}\frac{1 - \sum_{k=0}^{p-1} \exp(-\lambda_\delta(n))\lambda_\delta(n)^k/k!}{1 - e^{-\lambda_\delta(n)}} &= \frac{\lambda_\delta(n)^{p-1}}{p!} + \cdots \\ &= \frac{1}{p!}\left(\frac{n^2}{2N}\right)^{p-1} + \cdots\end{aligned}$$

The accumulated rate over the entire genealogy is

$$\begin{aligned}\Lambda_{2^p}(n) &= \frac{1}{p!(2N)^{p-1}} \int_0^n x^{2p-2}\, dx \\ &= \frac{n^{2p-1}}{(2p-1)p!(2N)^{p-1}}.\end{aligned}$$

Thus, the correct scaling for the onset of p-fold binary mergers is $n = \alpha N^{\frac{p-1}{2p-1}}$. The scaling was obtained in the previous chapter, but not the Poisson approximation.

3.2.3 Triple mergers

The reasoning for triple mergers is slightly different. We first need to obtain the rate of triple mergers during a single backward WF step. Consider the $m+1$st sample. Each of the first m samples has the same parent as the $m+1$st sample with probability $1/N$, a rare event. Thus, the number of samples out of the first m that have the same parent as the $m+1$st sample is Poisson with rate m/N. The probability that two of them have the same parent as $m+1$, causing a triple merger, is

$$\frac{1}{2!}\left(\frac{m}{N}\right)^2 e^{-m/N}$$

approximately. Therefore, the accumulated rate of triple mergers over a single generation is

$$\begin{aligned}\lambda_3(n) &= \frac{1}{2}\int_0^n \left(\frac{x}{N}\right)^2 e^{-x/N}\, dx \\ &= \frac{n^3 \exp(-n/N)}{4N^2} + \cdots\end{aligned}$$

Triple mergers are a rare event for $n \ll N^{2/3}$. However, when accumulating the rate of triple mergers over the entire genealogy, it is essential to account for the WF genealogy skipping some sample sizes.

The expected δ when the sample size is n, given that $\delta \geq 1$, is $\lambda_\delta(n)/\left(1-\exp(-\lambda_\delta(n)\right)$. Thus, for $m \leq n$, we take the probability that m is reached to be $\left(1-\exp(-\lambda_\delta(m))\right)/\lambda_\delta(m)$.

For the accumulated rate of triple mergers, we obtain

$$\Lambda_3(n) = \int_0^n \lambda_3(x) \frac{1 - \exp(-\lambda_\delta(x))}{\lambda_\delta(x)} dx.$$

We may set $n = \alpha N^{1/2}$ and then use the approximation $\lambda_\delta(n) = n^2/2N$ to obtain

$$\Lambda_3(n) = \frac{\alpha^2}{6} + \frac{e^{-\alpha^2/2} - 1}{3} \tag{3.5}$$

We may then use the Poisson distribution and approximate the expected number of triple mergers in the genealogy as $\Lambda_3(n)$ (see Figure 3.1) or calculate the probability of k triple mergers in the WF genealogy of the sample. For example, if $n = \alpha N^{1/2}$, the expected number of triple mergers in the genealogy is $\alpha^2/6 + \exp(-\alpha^2/2)/3 - 1/3$ in the limit of large N. The $N^{1/2}$ scaling of triple mergers was established in Chapter 2.

3.3 Perturbative analysis of the WF site frequency spectrum

The manner in which coalescent and WF genealogies differ may be inferred from (3.4), (3.5), and other similar results. Such differences in genealogy are a part of modeling and are not directly observable from sequence data. The question becomes to what extent the genealogical differences show up in sequence data.

In this section, we will outline the the main ideas in obtaining the WF frequency spectrum. The leading term of course is the coalescent answer, which is $1/j\mathcal{H}_{n-1}$. We will calculate the following N^{-1} terms.

The first perturbing terms, which we will calculate, suggests that the correct scaling for the divergence of WF frequency spectrum from that of the coalescent is $n = \alpha N$. Although not a proof, the suggestion is almost surely correct and we verify it from another angle later. The scaling for the onset of simultaneous binary mergers and triple mergers is $N^{1/3}$ and $N^{1/2}$ (from (3.4) and (3.5)). The fact that

the divergence in the frequency spectrum sets in for much larger samples means that the frequency spectrum is not very sensitive to multiple mergers in the genealogy.

If the WF genealogy of a sample of size n progresses through sample sizes as in

$$n \to n-1 \to \cdots \to 2 \to 1$$

without skipping any sample size in-between n and 1, we denote that no-skip event by $\mathcal{S}_0$. If the WF genealogy skips from a sample size of $m+2$ to m, omitting $m+1$, we denote such a skip-to-m event by $\mathcal{S}_m$ for $m = n-2, \ldots, 1$. The sample size of $m+1$ is the only omission in $\mathcal{S}_m$.

Other patterns of skipping are possible. However, the probability of such events is $\mathcal{O}(N^{-2})$. For an $\mathcal{O}(N^{-1})$ calculation, we only need to consider $\mathcal{S}_0$ and $\mathcal{S}_m$.

The WF frequency spectrum is calculated under the assumption of exactly one mutation in the genealogy of the sample. Therefore, we define the event $\mathcal{S}_0^\mu$ to be $\mathcal{S}_0$ and exactly one mutation in the genealogy of the sample. The event $\mathcal{S}_m^\mu$ is defined analogously.

3.3.1 Coalescent and WF propagators

The general approach to derive the WF frequency spectrum is to first determine the probability that a mutation occurs in the genealogy when the sample size is m for $m = n, \ldots, 2$. That would mean that 1 out of m ancestral samples is a mutant at some point in the genealogy. That probability is then propagated to the current sample size of n. We begin by studying propagation under the coalescent.

Suppose an ancestral sample of size m has $i \geq 1$ mutants and $(m-i)$ non-mutants. The genealogy from the current sample of size n to the ancestral sample of size m is assumed to involve only binary mergers, with no mutations in-between. As we will presently show, the probability that the current sample has $j \geq i$ mutants is given

by

$$\frac{(j-1)^{\underline{(i-1)}}}{(i-1)!}.\frac{(m-1)^{\underline{i}}(n-m)^{\underline{j-i}}}{(n-1)^{\underline{j}}}, \tag{3.6}$$

where $j^{\underline{i}}$ is the falling power $j(j-1)\dots(j-i+1)$ [24]. We adopt the convention that $j^{\underline{0}}$ is 1, even for $j=0$.

The Kingman partition distribution may be used to obtain (3.6). However, we will give a more direct argument. Suppose an ancestral sample of size $\mathfrak{m}$ has $\mathfrak{i}$ mutants. Suppose that an ancestral sample of size $\mathfrak{m}+1$ is related to it through a single binary merger. Then the probability that the sample of $\mathfrak{m}+1$ has $\mathfrak{i}+1$ mutants is $\mathfrak{i}/\mathfrak{m}$ because each sample out of $\mathfrak{m}$ is equally like to "split" and one of the $\mathfrak{i}$ mutants will split with probability $\mathfrak{i}/\mathfrak{m}$. Similarly, the probability that the sample of $\mathfrak{m}+1$ has $\mathfrak{i}$ mutants is $(\mathfrak{m}-\mathfrak{i})/\mathfrak{m}$ (in this case, one of the $\mathfrak{m}-\mathfrak{i}$ non-mutants has to split).

From here, we can write down the probability that a sample of n has j mutants when it is descended through binary splits from a sample of size m with i mutants to be

$$\binom{n-m}{j-i}\frac{((m-i)\dots(m-j-1))\,(i\dots(j-1))}{m\dots n-1}.$$

The argument for this expression is as follows. There are $n-m$ splits from n to m. The binomial coefficient chooses $j-i$ of those splits to be ones that increase the number of mutants. The denominator of the fraction in the expression steps from the sample size of m to the sample size of $n-1$ because those are the sample sizes that split. The numerator has the factor $(m-i)\dots(m-j-1)$ to account for splits of non-mutants. The other factor $i\dots(j-1)$ accounts for splits of mutants. The above expression is simplified to obtain (3.6).

When $i=1$, (3.6) reduces to

$$\frac{(n-m)^{\underline{j-1}}}{(n-1)^{\underline{j}}}(m-1), \tag{3.7}$$

a useful special case. Setting $j = 1$, we find the probability of a single mutant in the sample of n given a single mutant in the ancestral sample to be

$$\frac{m-1}{n-1} \tag{3.8}$$

which is another useful special case.

3.3.2 WF propagators

Suppose next that a sample of size n is descended from a sample of size m with i mutants through a single backward WF step. It is assumed that there are no mutations during this descent. The probability of j mutants in the current sample is then given by

$$\binom{n}{j}\left(\left\{{j \atop i}\right\} i!\right)\left(\left\{{n-j \atop m-i}\right\}(m-i)!\right) \Big/ \left\{{n \atop m}\right\} m!. \tag{3.9}$$

That is because in the current sample, we can choose j individuals to be mutants in $\binom{n}{j}$ ways. That being done, the j mutants in the current sample can be assigned to i mutants in the parental sample, with each parent receiving at least one child, in $\left\{{j \atop i}\right\} i!$ ways: the j samples can be partitioned into i in $\left\{{j \atop i}\right\}$ways and then can be permuted in $i!$ ways. The last bracketed factor in the numerator is the number of ways to assign $(n-j)$ not mutants to $(m-i)$ non-mutants in the parental sample. The denominator is the number of ways to assign n children to m parents, with each parent receiving at least one child.

The Stirling numbers (of the second kind) $\left\{{\mathfrak{n} \atop 1}\right\}$, $\left\{{\mathfrak{n} \atop \mathfrak{n}-1}\right\}$, and $\left\{{\mathfrak{n} \atop \mathfrak{n}-2}\right\}$ are given by 1, $\mathfrak{n}(\mathfrak{n}-1)/2$, and $\mathfrak{n}(\mathfrak{n}-1)(\mathfrak{n}-2)(3\mathfrak{n}-5)/24$, respectively [24]. Using (3.9) along with those formulas, we obtain the probabilities that a sample of size $m+2$ has $i, i+1, i+2$ mutants when it is descended from an ancestral sample of size m in a

single WF generation to be

$$\frac{(m-i)(m-i+1)}{m(m+1)} - \frac{2i(m-i)}{m(m+1)(3m+1)}, \tag{3.10a}$$

$$\frac{2i(m-i)}{m(m+1)} + \frac{4i(m-i)}{m(m+1)(3m+1)}, \tag{3.10b}$$

$$\frac{i(i+1)}{m(m+1)} - \frac{2i(m-i)}{m(m+1)(3m+1)}, \tag{3.10c}$$

respectively.

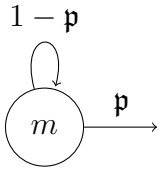
Suppose that a sample of size $m+2$ changes into a parental sample of size m under a single backward WF step. Given that the parental sample of m has only a single mutant, the probability that the sample of size $m+2$ has only a single mutant is

$$\frac{m-1}{m+1} - \frac{2(m-1)}{m(m+1)(3m+1)}, \tag{3.11}$$

which is obtained by setting $i=1$ in (3.10a). Comparing against (3.8), we find that skipping a step under WF reduces the factor that propagates the probability of a single mutant.

3.3.3 Probability of mutation at m

What is the probability of a mutation event at m assuming that the sample size m is visited? Consider the following picture:



The picture is showing that an ancestral sample size of m remains m under a backward WF step with probability $1-\mathfrak{p}$ and exits to a lower sample size with probability

$\mathfrak{p}$. Neglecting μ^2 terms, the probability that a sample of size m will be hit with a mutation is

$$\sum_{k=0}^{\infty} k(m\mu)\mathfrak{p}(1-\mathfrak{p})^k,$$

where k is the number of returns from m to m.

Thus, with μ^2 terms neglected, the probability of being hit with a mutation at m is equal to $m\mu/\mathfrak{p}$. We may take

$$\begin{aligned}\mathfrak{p} &= 1 - \prod_{k=1}^{m-1}\left(1-\frac{k}{N}\right) \\ &= \frac{m(m-1)}{2N} - \frac{m(m-1)(m-2)(3m-1)}{24N^2} + \cdots\end{aligned}$$

by ignoring terms after N^{-2}. We get the probability of being hit with a mutation at m to be

$$(2N\mu)\left(\frac{1}{m-1} + \frac{(m-2)(3m-1)}{12N(m-1)} + \mathcal{O}(N^{-2})\right) + \mathcal{O}(\mu^2).$$

Neglecting μ^2 and N^{-2} terms, we denote the probability of being hit with a mutation at m by

$$(2N\mu)\left(\frac{1}{m-1} + \frac{\mu_m}{12N}\right), \tag{3.12}$$

where $\mu_m = (m-2)(3m-1)/(m-1) + \mathcal{O}(N^{-1})$.

In fact, because we are neglecting μ^2 terms, (3.12) gives the probability that there is a single mutation in the entire genealogy with that mutation occurring when the ancestral sample size is m.

3.3.4 Probability that $m+2$ skips to m

Suppose the ancestral sample size is $m+2$. What is the probability that the ancestral sample size skips over $m+1$ and goes directly to m under WF? The ancestral sample size could skip over both $m+1$ and m, but because we are neglecting N^{-2} terms, those possibilities may be ignored.

The probability that a backward WF applied to a sample of size $m+2$ results in a sample of size m, conditioned on a merger, is

$$\frac{\left(\frac{1}{N^2}\binom{m+2}{3}+\frac{3}{N^2}\binom{m+2}{4}\right)(1-1/N)\dots(1-(m-1)/N)}{1-(1-1/N)\dots(1-(m+1)/N)}.$$

There are $\binom{m+2}{3}$ possible triple mergers and $3\binom{m+2}{4}$ possible double binary mergers. The first factor in the numerator accounts for the probabilities of those. In both a triple merger and a double binary merger, a total of m parents must be chosen distinctly, which occurs with probability $(1-1/N)\dots(1-(m-1)/N)$. That accounts for the second factor in the numerator. The denominator is the probability that the number of parents of $m+2$ samples is fewer than $m+2$ in a single backward WF step.

Simplifying the above expression, we obtain the probability of skipping to m is

$$\frac{m(3m+1)}{12N}, \tag{3.13}$$

with N^{-2} terms neglected. If $s_m = m(3m+1)$, this probability can be taken to be $s_m/12N$.

3.3.5 The event $\mathcal{S}_0^\mu$ and $\mathbb{P}(j|\mathcal{S}_0^\mu)$

From (3.13), it follows that that the probability of $\mathcal{S}_0$, which visits each ancestral sample size in $\{1,\dots n\}$ is $\Pi_{m=1}^{n-2}(1-s_m/12N)$. Using (3.12), the probability of a single mutation in the genealogy is $(2N\mu)\left(\sum_{m=2}^{n}(1/(m-1)+\mu_m/12N\right)$, with μ^2 terms ignored and with N^{-2} terms ignored in the coefficient of $2N\mu$. The summation over $m=2,\dots,n$ sums over the probability of the single mutation occurring when the ancestral sample size is one of $2,\dots,n$.

Thus, the probability of $\mathcal{S}_0^\mu$ is

$$\Pi_{m=1}^{n-2}(1-s_m/12N)\times(2N\mu)\left(\sum_{m=2}^{n}(1/(m-1)+\mu_m/12N\right).$$

Simplifying and omitting N^{-2} terms in the coefficient of $2N\mu$, we get $\mathbb{P}(\mathcal{S}_0^\mu) = (2N\mu)W_0 + \mathcal{O}(\mu^2)$ with

$$\begin{aligned} W_0 &= \mathcal{H}_{n-1} - \frac{\mathcal{H}_{n-1}}{12N}\sum_{m=1}^{n-2} m(3m+1) + \frac{1}{12N}\sum_{m=2}^{n}\mu_m \\ &= \mathcal{H}_{n-1} - \frac{\mathcal{H}_{n-1}n(n^2-4n+5)}{12N} + \frac{(n-1)(3n-2)}{24N} \end{aligned} \tag{3.14}$$

and with N^{-2} terms neglected in W_0. The last step in (3.14) is gotten after a routine simplification. At this point, we can think of $\mathbb{P}(\mathcal{S}_0^\mu)$ as proportional to the weight W_0.

Let $\mathcal{M}_m$ be the event that a mutation occurs in the genealogy of the sample of size n when the ancestral sample size is m. From (3.12), we know that the probability that a mutation occurs at m but nowhere else in the genealogy is proportional to $1/(m-1) + \mu_m/12N$. Therefore,

$$\mathbb{P}\left(\mathcal{M}_m \middle| \mathcal{S}_0^\mu\right) = \frac{\frac{1}{m-1} + \frac{\mu_m}{12N}}{\mathcal{H}_{n-1} + \frac{1}{12N}\sum_{m=2}^{n}\mu_m},$$

where the denominator is obtained by summing over $m = 2, \ldots, n$. The right hand side above can be simplified to obtain

$$\mathbb{P}\left(\mathcal{M}_m \middle| \mathcal{S}_0^\mu\right) = \frac{1}{(m-1)\mathcal{H}_{n-1}} + \frac{3m-4}{12N\mathcal{H}_{n-1}} - \frac{(n-1)(3n-2)}{24N\mathcal{H}_{n-1}^2(m-1)} + \cdots$$

with N^{-2} terms ignored and in the limit $\mu \to 0$.

The number of mutants in the current sample of size n is always denoted by j. The next step is to calculate $\mathbb{P}(j \big| \mathcal{M}_m, \mathcal{S}_0^\mu)$. For $j = 1$, we can use (3.8) to propagate a single mutant from an ancestral sample of size m to the current sample of size n and get

$$\mathbb{P}\left(j = 1 \middle| \mathcal{M}_m, \mathcal{S}_0^\mu\right) = \frac{m-1}{n-1}.$$

More generally, the probability $\mathbb{P}(j \big| \mathcal{M}_m, \mathcal{S}_0^\mu)$, where j stands for j mutants in the current sample of n, is given by (3.7). By writing $(3m-4)(m-1)$ as $3(m-1)^2-(m-1)$,

we obtain

$$\mathbb{P}(j\big|\mathcal{M}_m,\mathcal{S}_0^\mu)\mathbb{P}(\mathcal{M}_m\big|\mathcal{S}_0^\mu)=\frac{(n-m)^{\underline{j-1}}}{\mathcal{H}_{n-1}(n-1)^{\underline{j}}}+\frac{(m-1)^2(n-m)^{\underline{j-1}}}{4N\mathcal{H}_{n-1}(n-1)^{\underline{j}}}-\frac{(m-1)(n-m)^{\underline{j-1}}}{12N\mathcal{H}_{n-1}(n-1)^{\underline{j}}}$$
$$-\frac{(n-1)(3n-2)(n-m)^{\underline{j-1}}}{24N\mathcal{H}_{n-1}^2(n-1)^{\underline{j}}},$$

with N^{-2} terms ignored and in the limit $\mu\to 0$. We then have

$$\begin{aligned}\mathbb{P}(j\big|\mathcal{S}_0^\mu)&=\sum_{m=2}^{n}\mathbb{P}(j\big|\mathcal{M}_m,\mathcal{S}_0^\mu)\mathbb{P}(\mathcal{M}_m\big|\mathcal{S}_0^\mu)\\&=\frac{1}{\mathcal{H}_{n-1}j}+\frac{n(2n-j)}{j(j+1)(j+2)}-\frac{n}{12N\mathcal{H}_{n-1}j(j+1)}-\frac{(n-1)(3n-2)}{24N\mathcal{H}_{n-1}^2j}\end{aligned}$$

after simplification, with N^{-2} terms ignored and in the limit $\mu\to 0$. The simplification is effected using the following identities:

$$\begin{aligned}\sum_{m=2}^{n}(n-m)^{\underline{j-1}}&=(n-1)^{\underline{j}}/j,\\\sum_{m=2}^{n}(m-1)(n-m)^{\underline{j-1}}&=n(n-1)^{\underline{j}}/j(j+1),\\\sum_{m=2}^{n}(m-1)^2(n-m)^{\underline{j-1}}&=n(2n-j)(n-1)^{\underline{j}}/j(j+1)(j+2),\end{aligned}$$

for $j=1,2,\ldots$, each of which is easily proved by induction on n. Another method of proof is to begin with the difference identity $(n+1)^{\underline{j}}-n^{\underline{j}}=jn^{\underline{j}}$.

3.3.6 The event $\mathcal{S}_m^\mu$ and $\mathbb{P}(j|\,(\mathcal{S}_m^\mu)$

From (3.13), $\mathbb{P}(S_m)$, which is the probability the genealogy skips from sample size $m+2$ to m, is

$$\frac{m(3m+1)}{12N}\times\prod_{\ell\in\{1\ldots n-2\}-\{m,m+1\}}\left(1-\frac{\ell(3\ell+1)}{12N}\right)$$

or simply $m(3m+1)/12N$ with N^{-2} terms neglected.

Because $\mathbb{P}(S_m)$ leads with a N^{-1} term, we may simplify (3.12) and take the probability that a mutation hits when the ancestral sample size is ℓ to be $(2N\mu)/(\ell-$

1). It follows that $\mathbb{P}\left(\mathcal{S}_m^\mu\right) = (2N\mu)W_m + \mathcal{O}(\mu^2)$

$$W_m = \frac{m(3m+1)}{12N}\left(\mathcal{H}_{n-1} - \frac{1}{m}\right). \tag{3.15}$$

with N^{-2} terms neglected in W_m. At this point, we can take $\mathbb{P}(\mathcal{S}_m^\mu)$ to be proportional to W_m.

To calculate $\mathbb{P}(j\big|\mathcal{S}_m^\mu)$, we use a shortcut that greatly simplifies the algebra. Under the condition $\mathcal{S}_m^\mu$ and by (3.12) (with the $\mu_m/12N$ term ignored because W_m leads with a N^{-1} term), the probability of a mutation at ℓ is proportional to $1/(\ell-1)$ for $\ell \in \{2,\ldots,n\} - \{m+1\}$. Therefore the probability of a mutation at ℓ under the condition $\mathcal{S}_m^\mu$ is equal to

$$\frac{1/(\ell-1)}{\mathcal{H}_{n-1} - 1/m},$$

in the limit $\mu \to 0$ and with N^{-1} terms ignored. Now for the shortcut, suppose we can ignore the WF corrections to the propagators, namely, the latter terms in the WF propagators (3.10a), (3.10b), and (3.10c). ,We can then obtain the probability of j mutants in the current sample of n to be

$$\frac{1}{\mathcal{H}_{n-1} - 1/m} \sum_{\ell \in \{2,\ldots,n\} - \{m+1\}} \frac{1}{(\ell-1)} \times \frac{(\ell-1)(n-\ell)^{\underline{j-1}}}{(n-1)^{\underline{j}}},$$

where the single mutant at ℓ is propagated to n using the coalescent propagator (3.7) before summing over ℓ. This expression can be simplified to get

$$\frac{1}{\mathcal{H}_{n-1} - 1/m}\left(\frac{1}{j} - \frac{(n-m-1)^{\underline{j-1}}}{(n-1)^{\underline{j}}}\right), \tag{3.16}$$

which is the probability of j mutants except for the corrections given by the latter terms in the WF propagators (3.10a), (3.10b), and (3.10c).

We will now calculate the corrections separately. Let $\mathcal{M}_{2\ldots m}$ denote $\mathcal{M}_2 \cup \ldots \cup \mathcal{M}_m$, in words, the event where a mutation occurs when the ancestral sample size

is $2, \ldots, m$. The probability that a mutation strikes when the sample size is ℓ is proportional to $1/(\ell - 1)$. Therefore,

$$\mathbb{P}(\mathcal{M}_{2\ldots m} \big| \mathcal{S}_m^\mu) = \frac{\mathcal{H}_{m-1}}{(\mathcal{H}_{n-1} - 1/m)},$$

with all N^{-1} and μ terms ignored. The latter terms in the WF propagators (3.10a), (3.10b), and (3.10c) will be activated only when the condition $\mathcal{M}_{2\ldots m}$ holds in addition to $\mathcal{S}_m^\mu$.

Conditioning on $\mathcal{M}_{2\ldots m}$ and $\mathcal{S}_m^\mu$, the frequency spectrum of ancestral sample of size m is given by

$$1/i\mathcal{H}_{m-1}$$

for the probability of i mutants, $i = 1, \ldots, m-1$ (in the limit $\mu \to 0$ and with N^{-1} terms neglected). To obtain the correction, this frequency spectrum must first be propagated to $m+2$ samples using the latter terms of the WF propagators (3.10a), (3.10b), and (3.10c) because the condition $\mathcal{S}_m^\mu$ stipulates a skip form sample size $m+2$ to sample size m. Propagating the probabilities to $m+2$, we get the corrections to the probability of i mutants in a sample of $m+2$ under the conditions $\mathcal{M}_{2\ldots m}$ and $\mathcal{S}_m^\mu$ to be

$$\begin{aligned}
&\frac{-2(m-1)}{\mathcal{H}_{m-1} m(m+1)(3m+1)} \text{ for } i = 1,\\
&\frac{2m}{\mathcal{H}_{m-1} m(m+1)(3m+1)} \text{ for } i = 2,\\
&\frac{-2}{\mathcal{H}_{m-1} m(m+1)(3m+1)} \text{ for } i = m+1,
\end{aligned}$$

and zero for all other $i \in \{1, \ldots, m+1\} - \{1, 2, m+1\}$. Multiplying these numbers with the coalescent propagator (3.6) with $m \leftarrow m+2$ and $i \leftarrow 1, 2, m+1$, respectively, we get the corrections to the probability of j mutants in the current sample of n under the conditions $\mathcal{M}_{2\ldots m}$ and $\mathcal{S}_m^\mu$ to be

$$\frac{-2(m-1)(n-m-2)^{\underline{j-1}}}{\mathcal{H}_{m-1}m(3m+1)(n-1)^{\underline{j}}},$$
$$\frac{2(j-1)m(n-m-2)^{\underline{j-2}}}{\mathcal{H}_{m-1}(3m+1)(n-1)^{\underline{j}}},$$
$$\frac{-2(j-1)^{\underline{m}}(n-m-2)^{\underline{j-m-1}}}{\mathcal{H}_{m-1}m(3m+1)(n-1)^{\underline{j}}}.$$

Multiplying these terms by $\mathbb{P}(\mathcal{M}_{2\dots m}\big|\mathcal{S}_m^\mu)$ and adding to (3.16), we get

$$\begin{aligned}\mathbb{P}(j\big|\mathcal{S}_m^\mu) =&\frac{1}{\mathcal{H}_{n-1}-1/m}\left(\frac{1}{j}-\frac{(n-m-1)^{\underline{j-1}}}{(n-1)^{\underline{j}}}\right)-\frac{2(m-1)(n-m-2)^{\underline{j-1}}}{(\mathcal{H}_{n-1}-1/m)m(3m+1)(n-1)^{\underline{j}}}\\&+\frac{2(j-1)m(n-m-2)^{\underline{j-2}}}{(\mathcal{H}_{n-1}-1/m)(3m+1)(n-1)^{\underline{j}}}\\&-\frac{2(j-1)^{\underline{m}}(n-m-2)^{\underline{j-m-1}}}{\mathcal{H}_{m-1}m(3m+1)(n-1)^{\underline{j}}},\end{aligned}$$

in the limit $\mu\to 0$ and with N^{-1} terms ignored.

3.3.7 WF sample frequency spectrum

The sum $\sum_{m=1}^{n-2}W_m\mathbb{P}(j\big|\mathcal{S}_m^\mu)$ may be simplified to get

$$\begin{aligned}&\frac{(n-2)(n-1)^2}{12Nj}+\frac{(3n-2)[j=1]}{12N}-\frac{n}{12Nj(j+1)}-\frac{n(2n-j)}{4Nj(j+1)(j+2)}\\&-\frac{(n-j-2)(n-j-1)}{6Nj(j+1)(n-1)}+\frac{(2n-j-1)[j\geq 2]}{6Nj(j+1)}-\frac{(j-1)}{6N(n-1)(n-j)},\end{aligned}\tag{3.17}$$

where the second line accounts for WF corrections to the coalescent propagators. The simplification uses the identities

$$\sum_{m=1}^{n-2}(n-m-2)^{\underline{j-1}}=(n-2)^{\underline{j}}\Big/j \qquad \text{for } j=1,2,\dots$$
$$\sum_{m=1}^{n-2}(m-1)(n-m-2)^{\underline{j-1}}=(n-2)^{\underline{j+1}}\Big/j(j+1) \qquad \text{for } j=1,2,\dots$$
$$\sum_{m=1}^{n-2}m^2(n-m-2)^{\underline{j-2}}=(2n-j-1)(n-1)^{\underline{j}}\Big/(j-1)j(j+1) \qquad \text{for } j=2,3,\dots$$
$$\sum_{m=1}^{n-2}(j-1)^{\underline{m}}(n-m-2)^{\underline{j-m-1}}=(j-1)(n-2)^{\underline{j-2}} \qquad \text{for } j=1,2,\dots$$

In the last identity, $\mathfrak{a}^{\underline{\mathfrak{b}}}$ is assumed to be 1 if $\mathfrak{b}\leq 0$. All these identities may be verified by induction on n.

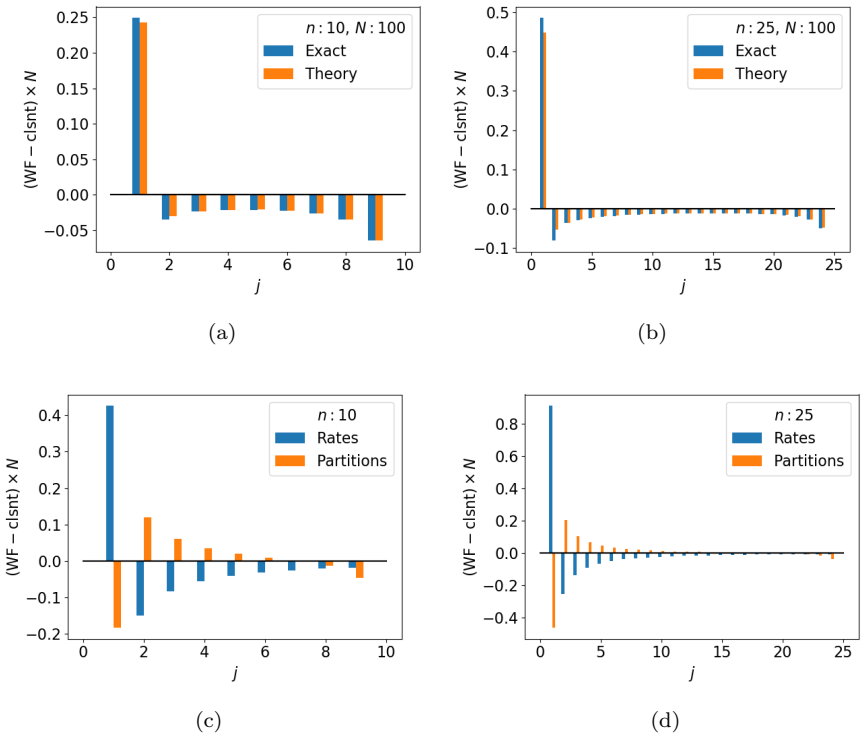


Figure 3.2: (a) and (b): WF minus coalescent computed using (3.2) minus (3.1) (theory) is compared with a computation using the program of [5] (exact). (c) and (d): (3.2) minus (3.1) minus (3.19) (rates) is compared with (3.19) (partitions).

The WF sample frequency spectrum (3.2) is obtained by simplifying

$$\frac{W_0\mathbb{P}(j\big|\mathcal{S}_0^\mu)+\sum_{m=1}^{n-2}W_m\mathbb{P}(j\big|\mathcal{S}_m^\mu)}{W_0+\sum_{m=1}^{n-2}W_m}. \tag{3.18}$$

If we look at the sequence steps building up to this point, the difference in the way WF and the coalescent partition children between parents first comes up in the latter term of (3.11) as well as (3.10a), (3.10b), (3.10c). That terms propagates to the second line of (3.17).

Thus the N^{-1} terms in the WF frequency spectrum (3.2) due to differences in partitioning between WF and the coalescent are given by

$$-\frac{(n-j-2)(n-j-1)}{6N\mathcal{H}_{n-1}(n-1)j(j+1)}+\frac{(2n-j-1)[j\geq 2]}{6N\mathcal{H}_{n-1}j(j+1)}-\frac{j-1}{6N\mathcal{H}_{n-1}(n-1)(n-j)}. \tag{3.19}$$

Evaluating with $j=1$ and retaining only the dominant term, we get $-n/12N\mathcal{H}_{n-1}$

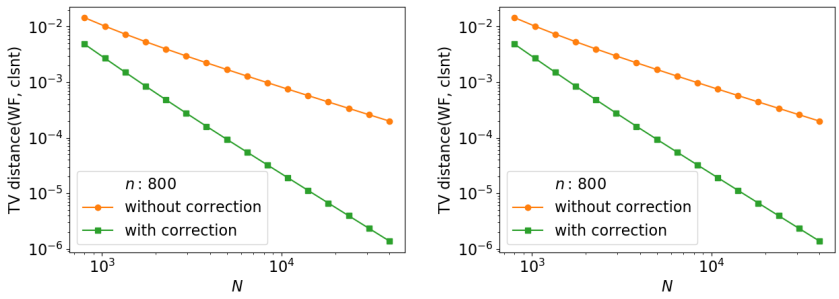


Figure 3.3: Total variation distance between WF frequency spectrum [5] and that of the coalescent given by (3.1) (without correction) or with correction as given by (3.2).

to be the effect on singleton probability of the difference in partitioning distributions. Evaluating (3.2) with $j = 1$ and retaining only the dominant term, we obtain $n/12N\mathcal{H}_{n-1}$ as the amount by which the WF singleton probability exceeds that of the coalescent. Therefore, the effect of the mismatch in rates of merger (as defined in the Introduction) must be $n/6N\mathcal{H}_{n-1}$.

Figure 3.2 shows that the WF singleton probabilities are elevated and the rest of the frequency spectrum is depressed, as may be inferred from the last two terms of (3.2). The figure also illustrates the correction due to rates being twice as high as the correction due to differences in the way children are partitioned between parents.

Because $j = 1$ singleton probabilities are elevated under WF and other probabilities are lowered, we may obtain the total variation distance between WF and the coalescent by simply taking the difference in $j = 1$ probabilities. Thus, the perturbative estimate for the total variation distance between the WF frequency spectrum and that of the coalescent is $n/12N\mathcal{H}_{n-1}$. This estimate is qualitatively correct even for $n = N$, and even quantitatively it is not unreasonable, being about 2/3rds of a better estimate we will presently derive. Figure 3.3 shows that the total variation distance increases with n and decreases with N.

If the number of samples is $n = 3$, the exact WF frequency spectrum is given by

$$\frac{2N-1}{3N-2}, \quad \frac{N-1}{3N-2}.$$

If $n = 4$, the exact WF frequency spectrum is given by

$$\frac{2\left(9N^3-20N^2+16N-4\right)}{33N^3-82N^2+73N-22},$$
$$\frac{3\left(N^2-2N+1\right)}{11N^2-20N+11},$$
$$\frac{2\left(N-1\right)\left(3N^2-6N+4\right)}{\left(3N-2\right)\left(11N^2-20N+11\right)}.$$

The perturbative WF frequency spectrum (3.2) may be checked against these exact answers.

3.4 Population sized samples

Suppose the sample size is $n = \alpha N$. For an individual among the parental population of N, the probability that any given sample is a child is $1/N$, a rare event. The accumulated probability over the sample of size αN is α. Therefore, by the Poisson clumping heuristic, we may approximate the number of children of an individual in the parental generation by the Poisson distribution with rate α. The probability that an individual has k children among the αN samples is approximately $\exp(-\alpha)\alpha^k/k!$. The generating function $\sum_{k=0}^{\infty} \mathfrak{p}_k x^k$ with $\mathfrak{p}_k = \exp(-\alpha)\alpha^k/k!$ is $\exp(\alpha(x-1))$.

The only individuals in the parental generation that appear in the genealogy are ones who have at least one child among the samples. Therefore, it is natural to look at the Poisson distribution under the condition of having one child. Under that condition, the probability of having k children is $\mathfrak{p}_k/(1-\exp(-\alpha))$ and the generating function is $(\exp(\alpha x)-1)/(\exp(\alpha)-1)$.

Let $G_1(\alpha N) = g_1(\alpha)N$ be the expected b-branch length with $b = 1$ of the WF genealogy of a sample of size αN. By (3.3), the expected number of parents is

$N(1-\exp(-\alpha))$. Thus, we may write

$$g_1(\alpha)N = N\alpha + \mathfrak{f}Ng_1(1-\exp(-\alpha))$$

because the current samples $N\alpha$ all contribute to the 1-branch length and with the understanding that $\mathfrak{f}$ is the probability that a branch with a single descendant in the genealogy of the parental sample of size $(1-\exp(-\alpha))N$ remains a branch with a single descendant in the genealogy of the current sample of size αN. That probability $\mathfrak{f}$ is the same as the probability of a parent having a single child, which is $\alpha/(\exp(\alpha)-1)$. Therefore, we have

$$g_1(\alpha) = \alpha + \frac{\alpha}{\exp(\alpha)-1} g_1\left(1-\exp(-\alpha)\right), \tag{3.20}$$

which is a result of [92] derived essentially using their arguments.

The generating function for the number of children of a parents is approximately

$$\left(\frac{\exp(\alpha x)-1}{\exp(\alpha)-1}\right)^a. \tag{3.21}$$

Using [24, p. 265] to evaluate the sum, the probability that a parents have b children is found to be

$$\frac{\alpha^b a!}{(\exp(\alpha)-1)^a b!}\left\{{b \atop a}\right\} \tag{3.22}$$

for $b = a, a+1, \ldots$ Here $\left\{{b \atop a}\right\}$ is a Stirling number of the second kind [24]. Using the same argument as above and taking the b-branch length with αN samples to be $G_b(\alpha N) = Ng_b(\alpha)$, we get the recurrence

$$g_b(\alpha) = \sum_{a=1}^{b} g_a(1-\exp(-\alpha)) \times \frac{\alpha^b a!}{(\exp(\alpha)-1)^a b!}\left\{{b \atop a}\right\} \tag{3.23}$$

for $b = 2, 3, \ldots$

By solving the recurrences for $g_b(\alpha)$ and taking $\alpha = 1$, we can obtain approximations to the WF frequency spectrum with $n = N$ and compare it to (3.1), which is

the coalescent frequency spectrum. However, we seek to separate the difference into a part due to the mismatch in rates of mergers and a part due to the difference in the way children are partitioned among parents.

We turn to the discrete coalescent, which is a model intermediate between the coalescent and WF. To obtain the manner in which αN children are split between βN parents under the discrete coalescent, which uses the Kingman partition distribution, we may fix an orange at the left most position and permute $\beta N - 1$ identical oranges and $\alpha N - \beta N$ identical apples after it. The number of children of the ith parent can be taken to be the number of apples between the ith and $i+1$st orange plus one (thus counting the ith orange) [11, 29].

The probability that a parent has k children is approximately $\gamma(1-\gamma)^k$, with $\gamma = \beta/\alpha = (1-\exp(-\alpha))/\alpha$ for a sample of size αN. The generating function of this geometric distribution is $\gamma x/\left(1-(1-\gamma)x\right)$. The generating function for the number of children of a parents is approximately $\left(\gamma x/\left(1-(1-\gamma)x\right)\right)^a$. By extracting the coefficient of x^b, we find the probability of a parents having b children under the discrete coalescent to be

$$\binom{b-1}{a-1}\gamma^a(1-\gamma)^{b-a}$$

approximately.

If $\tilde{G}_b(\alpha N)$ denotes the b-branch length of the discrete coalescent genealogy of αN samples, we may set $\tilde{G}_b(\alpha N) = N\tilde{g}_b(\alpha)$ and obtain the recurrences

$$\begin{aligned}\tilde{g}_1(\alpha) &= \alpha + \frac{1-\exp(-\alpha)}{\alpha}\tilde{g}_1\left(1-\exp(-\alpha)\right)\\ \tilde{g}_b(\alpha) &= \sum_{a=1}^{b}\tilde{g}_a(1-\exp(-\alpha)) \times \binom{b-1}{a-1}\gamma^a(1-\gamma)^{b-a},\end{aligned} \tag{3.24}$$

where $b = 2, 3, \ldots$

b	ϵ_b	$\tilde{\epsilon}_b$
1	0.240917257	0.418035261
2	-0.046223840	-0.100136471
3	0.005196946	-0.032826669
4	0.001095702	-0.017086273
5	-0.000238278	-0.011181848
6	-0.000114882	-0.008036411
7	-0.000004091	-0.006053860

Table 3.1: The expected b-branch length of the WF genealogy of $n = N$ samples is $(2/b + \epsilon_b)N$. For the discrete coalescent, whose merger rates match WF but which partitions children between parents like the coalescent, it is $(2/b + \tilde{\epsilon}_b)N$.

In the Appendix, we show how to solve (3.20), (3.23), and (3.24) accurately using Chebyshev polynomials. For the coalescent, the expected b-branch length for a sample of size $n = N$ is $2N/b$. Therefore, we set $g_b(1) = (2/b + \epsilon_b)N$ and $\tilde{g}_b(1) = (2/b + \tilde{\epsilon}_b)N$ and report ϵ_b and $\tilde{\epsilon}_b$ in Table 3.1.

The first column of the table agrees very well with [17, p. 214]. To obtain the total size of the WF genealogy of $n = N$ samples, we use

$$\begin{aligned}\sum_{b=1}^{N-1} G_b(N) &= N\left(2\mathcal{H}_{N-1} + \sum_{b=1}^{N-1}\epsilon_b\right)\\ &= N\left(2\mathcal{H}_{N-1} + \Delta\right).\end{aligned}$$

We estimate Δ to be 0.200645075 by summing ϵ_b over $1 \le b \le 20$. The size of the discrete coalescent genealogy is the same as that of WF genealogy by definition. Our value for Δ agrees with Fisher's except in the last decimal place.

The probability of j mutants in the WF spectrum of $n = N$ samples is estimated to be

$$\frac{G_b(N)}{N(2\mathcal{H}_{N-1} + \Delta)} = \frac{1}{j\mathcal{H}_{N-1}} + \frac{\epsilon_j\mathcal{H}_{N-1} - \Delta/j}{\mathcal{H}_{N-1}\left(2\mathcal{H}_{N-1} + \Delta\right)}.$$

The estimated probability of $j = 1$ under WF exceeds $1/j\mathcal{H}_{N-1}$ because $\epsilon_1 > \Delta$. For $j = 3$, the term $\epsilon_j\mathcal{H}_{N-1} - \Delta/j$ is negative as long as $N < 6.8 \times 10^5$ but flips sign around $N = 6.8 \times 10^5$. For $j = 4$, $\epsilon_j\mathcal{H}_{N-1} - \Delta/j$ turns positive only around

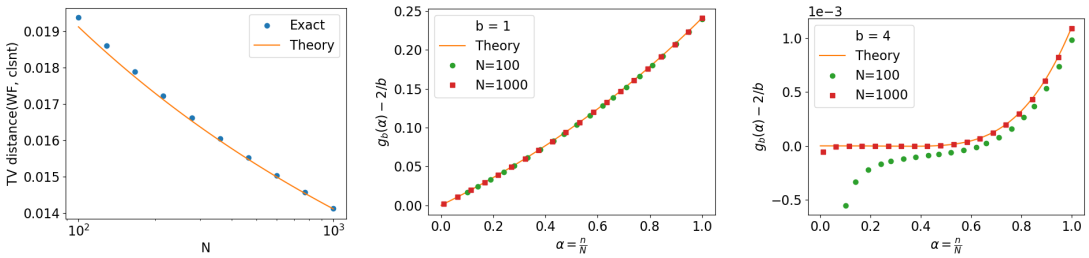


Figure 3.4: The first plot demonstrates the accuracy of (3.25). The next two plots examine the accuracy of $g_b(\alpha)$. In all cases, the exact computations use the computer program of [5].

$N = 10^{20}$. Thus, with minor caveats, the WF frequency spectrum is elevated at $j = 1$ but depressed slightly for $j > 1$.

We can approximate the total variation distance between WF and coalescent frequency spectrums for $n = N$ as

$$\frac{\epsilon_j \mathcal{H}_{N-1} - \Delta/j}{\mathcal{H}_{N-1}\left(2\mathcal{H}_{N-1} + \Delta\right)} = \frac{0.1204}{\log N} - \frac{0.1819}{(\log N)^2} + \cdots, \tag{3.25}$$

a result that is a direct consequence of [17]. The first plot of Figure 3.4 shows this estimate to be quite good. The figure also shows $g_1(\alpha)$ is quite accurate for even $N = 100$ and small α, although $g_4(\alpha)$ has visible errors for $N = 100$.

From Table 3.1, it is evident that the excess of the discrete coalescent's j mutant probability over that of the coalescent

$$\frac{\tilde{\epsilon}_j \mathcal{H}_{N-1} - \Delta/j}{\mathcal{H}_{N-1}\left(2\mathcal{H}_{N-1} + \Delta\right)}$$

is positive for $j = 1$ and negative for $j > 1$. The discrete coalescent and the coalescent differ only with respect to their rates of merger. Both of them follow the Kingman partition distribution. The effect due to difference in the rates of merger alone is twice as great because $\tilde{\epsilon}_1$ is nearly twice ϵ_1.

When $n \ll N^{1/2}$, we can say that the coalescent is faster than WF [20] because $n(n-1)/2N \geq \mathbb{E}\delta$ (see Appendix). However, when $n \gg N^{1/2}$, there can be several mergers in the same generation and rates of merger cannot be compared so directly.

Although, the coalescent begins with a higher rate it adjusts its rate downwards with every binary merger.

During a single backward WF step a sample size of $n = \alpha N$ changes on an average to $m = (1-\exp(-\alpha))N$. On an average the coalescent takes $2N(1/m-1/n)$ generations to go from n samples to m. In fact, $2((1-\exp(-\alpha))^{-1} - \alpha^{-1}) = 1 + \alpha/6 + \cdots > 1$ (see Appendix), and the coalescent is in fact slower.

However, the 1-branch length of the coalescent in going from n samples to m is equal to $2N(n-m)/(n-1)$. It may be shown that $2N(n-m)/(n-1) < N\alpha$ when $n = N\alpha$ and $m = N(1-\exp(-\alpha))$ (see Appendix). Therefore, although the coalescent may take a little more than a generation to go from n samples to m, its 1-branch length is lower as a consequence of repeated binary mergers over slightly more than a generation.

3.5 Discussion

WF deviates from the assumptions of the coalescent for even small sample sizes. Simultaneous binary mergers appear in WF genealogies for sample sizes of only $\alpha N^{1/3}$ with appreciable probability. Triple mergers appear for samples sizes of $\alpha N^{1/2}$.

However, the effect of such deviations on the site frequency spectrum is minimal. Deviations in the site frequency spectrum set in only for sample sizes αN. Even for population sized samples the deviation is only around 1%. The effect is so small because the coalescent is self-correcting. The rate of mergers under the coalescent is faster, but the coalescent lowers the rate with every merger. The coalescent limits itself to binary mergers. As a result, the offspring distribution under the coalescent is geometric, whereas it is Poisson under WF. The geometric and Poisson distributions are not far enough apart to cause a major effect. In addition, the effect of differing

offspring distributions partly cancels the effect of differing rates of merger.

Population substructure is perhaps the major reason to look for more sophisticated models than the coalescent [11, 89]. Skewed offspring distributions are another reason [13, 57]. In the setting of skewed offspring distributions, it is known that the skew has to be comparable to the population size for deviations to show up [13]. Thus, in that setting too, the coalescent is a robust model.

It is know that increasing skewness of offspring distribution raises the probability of singletons and lowers the probabilities of j mutants for $j > 1$ [13]. The Poisson offspring distribution of WF has a lower variance that the geometric offspring distribution of the coalescent (see Appendix). Our finding that the effect of differing offspring distributions is to lower the singleton probability under WF is consistent with this point of view.

As far as the single site frequency spectrum is concerned, the coalescent is a robust and reliable model relative to WF, and it will perhaps remain so until the SNP determination errors fall below a percent. However, what if multiple sites are considered, possibly allowing for recombination between sites? Our conjecture is that the total variation distance between WF and the coalescent will still be of the order $C/\log N$ for even population sized samples. However, the constant C may increase with the number of sites. In that regard, we mention the availability of software to efficiently simulate WF genealogies under very general conditions [70].

3.6 Appendix

In this appendix, we explain how to solve (3.20), (3.23), and (3.24) using Chebyshev polynomials. In addition, a few elementary inequalities used in the chapter are proved.

Chebyshev polynomials

For convenience, we restate the recurrence for g_1:

$$g_1(\alpha) = \alpha + \frac{\alpha}{\exp(\alpha) - 1} g_1 \left(1 - \exp(-\alpha)\right).$$

Begin with $g_1(\alpha) = C_0 + C_1\alpha + C_2\alpha^2 + C_3\alpha^3 + \cdots$ and expand each term of the recurrence to obtain all terms up to the α^3 term. We then obtain

$$C_0 + C_1\alpha + C_2\alpha^2 + C_3\alpha^3 = C_0 + (1 - C_0/2 + C_1)\,\alpha + (C_0/12 - C_1 + C_2)\,\alpha^2 + (C_1/2 - 3C_2/2 + C_3)\alpha^3.$$

If follows that $g_1(\alpha) = 2 + \alpha/6 + \alpha^2/18 + \cdots$ as in [92]. Using the same method, we get $g_2(\alpha) = 1 - \alpha^2/36 + \mathcal{O}(\alpha^3)$ and $g_b(\alpha) = 2/b + \mathcal{O}(\alpha^3)$ for $b > 2$.

To solve the recurrence for $g_1(\alpha)$, we set $g_1(\alpha) = 2 + \alpha/6 + \alpha^2/18 + \mathfrak{g}_1(\alpha)$. The resulting recurrence of $\mathfrak{g}_1(\alpha)$ is

$$\mathfrak{g}_1(\alpha) = \alpha - 2 - \alpha/6 - \alpha^2/18 + \frac{\alpha}{\exp(\alpha) - 1}(2 + \beta/6 + \beta^2/18 + \mathfrak{g}_1(\beta)),$$

where $\beta = 1 - \exp(-\alpha)$. It is solved by iteration at each of 32 Chebyshev points in $\alpha \in [0, 1]$. The function $g_1(\alpha)$ may then be obtained with 10+ digits of accuracy at any $\alpha \in [0, 1]$ using the barycentric Lagrange interpolant [87]. The functions $g_b(\alpha)$, with $b = 2, 3, \ldots, 20$ are calculated using the same method.

For the functions $\tilde{g}_b(\alpha)$, $b = 1, \ldots, 20$, we begin with $\tilde{g}_1(\alpha) = 2 + \alpha/3 + 2\alpha^2/27 + \tilde{\mathfrak{g}}_1(\alpha)$, $\tilde{g}_2(\alpha) = 1 - \alpha/18 - 19\alpha^2/432 + \tilde{\mathfrak{g}}_2(\alpha)$, and $\tilde{g}_b(\alpha) = 2/b - \alpha/3b(b+1) - \alpha^2/18b(b+1)(b+2) + \tilde{\mathfrak{g}}_b(\alpha)$ for $b > 2$. The rest of the method is the same.

The inequality $n(n-1)/2 \geq \mathbb{E}\delta$

To verify that $n(n-1)/2 \geq \mathbb{E}\delta = n - N + N(1 - 1/N)^n$, set

$$(1 - 1/N)^n = 1 - n/N + n(n-1)/2N^2 - \mathfrak{r}/6N^3$$

and use the Lagrange form of the Taylor series remainder to deduce $\mathfrak{r} > 0$.

The inequality $2((1-\exp(-\alpha))^{-1} - \alpha^{-1}) > 1$

The inequality $2((1-\exp(-\alpha))^{-1} - \alpha^{-1}) > 1$ is equivalent to

$$e^{\alpha} > \frac{e^{\alpha}-1}{2} + \frac{e^{\alpha}-1}{\alpha},$$

which is proved by verifying that the series for the left hand side majorizes the series for the right hand side.

The inequality $2N(n-m)/(n-1) < N\alpha$

To show that $2N(n-m)/(n-1) < N\alpha$ when $n = N\alpha$ and $m = N(1-\exp(-\alpha))$, first observe the inequality follows from $2(1-m/n) < \alpha$ for N large. Now $2(1-m/n) < \alpha$ is equivalent to $e^{-\alpha} < 1 - \alpha + \alpha^2/2$, which can be verified using the Lagrange form of the Taylor series remainder.

The inequality $\sigma_G > \sigma_P$

Suppose the sample size is αN with the parental population size being N as usual. Conditional on an individual of the parental generation being a parent of one of the samples and assuming N large, its number of children (among the samples) is given by the generating function $(\exp(\alpha x)-1)/(\exp(\alpha)-1)$. It follows that the expectation of the number of children is α and the variance is

$$\sigma_P = \frac{\alpha}{1-\exp(-\alpha)} + \frac{\alpha^2}{1-\exp -\alpha} - \frac{\alpha^2}{(1-\exp(-\alpha))^2}.$$

If the αN children are split among their parents according to the Kingman partition distribution, the generating function for the number of children is $\gamma x/(1-(1-\gamma)x$ with $\gamma = (1-\exp(-\alpha))/\alpha$. The expectation is again α and the variance is

$$\sigma_G = \frac{\alpha}{1-\exp(-\alpha)} + \frac{\alpha^2}{(1-\exp(-\alpha))^2} - 2.$$

One may verify that $\sigma_G > \sigma_P$ by plotting a graph. Alternatively,

$$\sigma_G - \sigma_P = \frac{\alpha^2}{(\exp(\alpha) - 1)^2} \left(2e^{\alpha} \frac{(e^{\alpha} + 1)}{2} - 2 \left(\frac{e^{\alpha} - 1}{\alpha} \right)^2 \right)$$

must be positive because the power series of both e^{α} and $(e^{\alpha} + 1)/2$ majorize the power series of $(e^{\alpha} - 1)/\alpha$.

Intuitively, we expect $\sigma_G > \sigma_P$ because the geometric distribution has exponential decay, whereas the Poisson distribution has super-exponential decay.

CHAPTER 4

The Site Frequency Spectrum under Finite and Time-Varying Mutation Rates[1]

4.1 Introduction

The mutation rate varies considerably across the human genome. CpG junctions are well-known to have particularly high mutation rates [12]. Some of the variation in the mutation rate across the human genome appears to be related to correlations between single nucleotide polymorphisms found in humans as well as other nearby primates [33, 37, 36, 42]. A variety of statistical models of the variation of the mutation rate have been proposed and examined [2, 7, 16, 36, 61]. The mutation rate can vary even between human populations [34, 56, 67].

The site frequency spectrum (SFS) is commonly used to summarize the effect of mutations across the genome. For a haploid sample of size n, the SFS consists of the probability that j of the samples carry the mutant allele, for $j = 1, \ldots, n-1$, at a polymorphic site. If the mutation rate itself varies widely across the genome, it is essential to know how the mutation rate affects the SFS. In this article, we derive an algorithm to calculate the SFS with mutation rates as well as population sizes allowed to vary in an arbitrary manner. The mutation rate is assumed to be $\mu(t)$ per base pair per generation at time t and the haploid population size is assumed to be $N(t)$. The algorithm relies on the coalescent approximation to genealogies [11].

[1] A modified version of this chapter, under the same title, is available on *biorxiv.org* [59]

In particular, n samples are assumed to experience a binary merger according to a Poisson process of rate $n(n-1)/2N(t)$. The samples are hit with mutations by an independent Poisson process of rate $\mu(t)n$.

An algorithm for calculating the SFS assuming μN to be negligible and μ to be constant in time is due to [74]. The Polanski-Kimmel algorithm,which relies on the earlier work of [29] as well as [73], is based on the internal structure of the coalescent genealogy. In particular, the algorithm relies on the expected branch length of the genealogy with exactly b descendants for $b = 1, \ldots, n-1$.

Our algorithm allows $\mu(t)$ to be finite and varying in time and is also based on the coalescent approximation. However, it pays no attention to the internal structure of the genealogy. The algorithm is more Markovian in spirit and is partly based on the ideas in the earlier analytic work described in Chapters 2 and 3.

[33] have presented data analysis showing that samples of size $n \approx 10^5$ have experienced more than one mutation at several polymorphic sites with $\mu \in [10^{-9}, 10^{-7}]$. Our algorithm can calculate the probability that a polymorphic site has experienced more than one mutation exactly. Using the demography inferred by [33], we precisely delineate the probability of more than one mutation in the genealogy.

When n and μ are large enough that a polymorphic site has been hit with either one, two, or more mutations, the SFS is a mixture of the SFS due to a single mutation and the SFS due to two or more mutations. Our calculations imply that the effects described by [33], such as the change in the profile of rare alleles at sites of higher μ, are mostly due to two or more mutations in the genealogy, which is in agreement with their conclusions.

Beginning with the work of [39], a number of authors have questioned the constancy of $\mu(t)$ with respect to t [47, 65, 64, 67, 79]. The germ line mutation rate

is known to depend on the number of cell divisions experienced by the germ line [23, **?**]. The number of cell divisions in the germ line is greater in the human male than the human female, and in the male it increases with age. The mutation rate in the male germ line is higher, as already deduced by [32] for the hemophilia gen. The dependence on the number of cell divisions could be due to errors during either genome replication or DNA repair [23], and some of the de novo mutations are shared between siblings in a mosaic pattern [76]. We use our algorithm to illustrate how increasing and decreasing mutation rates alter the SFS.

The coalescent and the diffusion equation often provide alternative routes to the same results. Accordingly, the SFS can be computed using the diffusion equation [26, 30, 78]. For the possibility of handling varying mutation rates using the diffusion equation, see [81, 88]. In the diffusion approach, the transition probabilities are first obtained and the SFS is computed using the transition probabilities. The sample size n enters only during the latter step. Therefore, it may appear as if the diffusion equation can calculate SFS for even large n with not much more trouble than for small n. However, for large n, the transition probabilities have to be calculated more accurately and with greater resolution.

4.2 Calculating the SFS under Finite and Varying Mutation Rates

Suppose the population size $N(t) \equiv N$ and mutation rate $\mu(t) \equiv \mu$ are both constant. If a site is polymorphic, the probability that j out of n samples carry the mutant allele converges to

$$\frac{1/j}{\mathcal{H}_{n-1}}, \tag{4.1}$$

where $\mathcal{H}_{n-1} = 1 + 1/2 + \cdots + 1/(n-1)$, in the limit $\mu N \to 0$ [11].

Let us now briefly examine the case where N and μ are constant, but without the

assumption of μN being negligible. The method we use to approach this problem will help clarify our reasoning behind the algorithm for the full case, where $N(t)$ and $\mu(t)$ both vary.

If the coalescent genealogy is sectioned at some fixed time in the past, we will refer to the lineages present at that time in the past as ancestral samples, following our earlier usage. Suppose the number of ancestral samples is k. Because the Poisson process of rate $k(k-1)/2N$ that produces a binary merger in the ancestral sample and the Poisson process of rate μk that hits the sample with a mutation are independent, the probability that the next event in the genealogy is a binary merger is

$$\frac{k(k-1)/2N}{k(k-1)/2N + 2k\mu} = \frac{k-1}{k-1+2N\mu}. \tag{4.2}$$

Correspondingly, the probability that the next event is a mutation is

$$\frac{2N\mu}{k-1+2N\mu}. \tag{4.3}$$

It follows that the probability $q_0(k)$ that k samples coalesce without being hit by a mutation is

$$q_0(k) = \prod_{j=2}^{k} \frac{j-1}{j-1+2N\mu}.$$

In more detail, for ancestral sample sizes $j = 2, \ldots, k$, a binary merger must precede a mutation, which occurs with a probability given by (4.2) (with $k \leftarrow j$) for each $j = 2, \ldots, k$.

Similarly, the probability $q_1(k)$ that a sample of size k coalesces after experiencing exactly one mutation is

$$q_1(k) = \prod_{j=2}^{k} \frac{j-1}{j-1+2N\mu} \times \sum_{j=2}^{k} \frac{2N\mu}{j-1+2N\mu}.$$

In more detail, the probability that a sample of size k coalesces after experiencing exactly one mutation when the ancestral sample size is j but with no other mutations

in the genealogy is

$$\prod_{\substack{\ell=2\\ \ell\neq j}}^{k}(\ell-1)/(\ell-1+2N\mu)\times\frac{2N\mu}{j-1+2N\mu}\times\frac{j-1}{j-1+2N\mu}.$$

The first factor occurs because the first event to hit an ancestral sample of size $\ell\neq j$ is a binary merger. The second factor occurs because the first event experienced by an ancestral sample of size j must be a mutation (whose probability is given by (4.3) with $k\leftarrow j$) and the third factor because the sample of size j then experiences a binary merger (whose probability is given by (4.2) with $k\leftarrow j$). The formula for $q_1(k)$ is obtained by summing over $j=2,\ldots,k$.

The condition or event that n samples coalesce with exactly one mutation in the genealogy will be denoted by $\mathcal{C}_n$. The definition of $\mathcal{C}_n$ will be changed slightly after we introduce the concept of an ancestral lens. Conditioned on $\mathcal{C}_n$, the probability that a mutation event occurs in the genealogy when the sample size is k is given by

$$\frac{\frac{1}{k-1+2N\mu}}{\sum_{j=2}^{n}\frac{1}{j-1+2N\mu}}.$$

Using the coalescent propagators derived in Chapter 3 and in [29], we may deduce the probability of j mutants in the sample under the condition $\mathcal{C}_n$ to be

$$\frac{1}{\sum_{k=2}^{n}\frac{1}{k-1+2N\mu}}\sum_{k=2}^{n}\frac{1}{k-1+2N\mu}\times\frac{(n-k)^{\underline{j-1}}(k-1)}{(n-1)^{\underline{j}}},\tag{4.4}$$

where $\mathfrak{a}^{\underline{\mathfrak{b}}}$ denotes the falling power $\mathfrak{a}(\mathfrak{a}-1)\ldots(\mathfrak{a}-\mathfrak{b}+1)$. The SFS (3.1) is obtained by substituting $\mu=0$ in this formula.

There is no obvious way to evaluate this formula for $j=1,\ldots,n-1$ in $\mathcal{O}(n)$ arithmetic operations for $\mu>0$, although the $\mu=0$ case given by (3.1) can be evaluated in $\mathcal{O}(n)$ arithmetic operations. The formula (4.4) for the SFS under the condition $\mathcal{C}_n$ can be cast in the form of a recurrence and the SFS evaluated using $\mathcal{O}(n^2)$ operations as seen in Chapter 2.

The derivation of the SFS conditioned on $\mathcal{C}_n$ for constant N and μ relies on $q_0(k)$, the probability that k samples coalesce with zero mutations, and $q_1(k)$, the probability that k samples coalesce with exactly one mutation in their genealogy. To apply a similar approach to the case with time varying $N(t)$ and $\mu(t)$, we will of course need to extend those functions to $q_0(k,t)$, the probability that k samples at time t coalesce with zero mutations, and $q_1(k,t)$, the probability that k samples at time t coalesce with exactly one mutation in their genealogy.

We will construct differential equations which we can use to solve for $q_0(k,t)$ and $q_1(k,t)$ and then construct another differential equation which, given q_0 and q_1 can be used to obtain the SFS. Before we discuss these equations, we must first deal with 2 issues: First, the range of t is infinite, so we must determine the range of useful values of t for determining the SFS to sufficient precision. Second, at any given t, the range of likely k is much smaller than $\{n,\ldots,1\}$. By restricting our computation to useful k, we will obtain significant savings in arithmetic operations and memory usage. We now turn to the concept of the ancestral lens, which allows us to deal with both of these issues.

4.2.1 The ancestral and calendar lens

We denote ancestral time by τ (generations), with the current epoch being $\tau = 0$. If the number of samples is n, let $p(k,\tau)$ be the probability that the number of ancestral samples at time τ is k. The ancestral lens is defined as the set of all (k,τ) such that $p(k,\tau) \geq \epsilon_{lens}$. The tolerance ϵ_{lens} is taken to be 10^{-40}. All samples sizes outside the ancestral lens have a negligible probability and may be ignored without affecting calculations for the current sample at $\tau = 0$. Additionally, the lens allows us to find a maximum τ, after which we do not need to calculate q_0 or q_1.

The probability that the ancestral sample is of size k at time τ and undergoes a

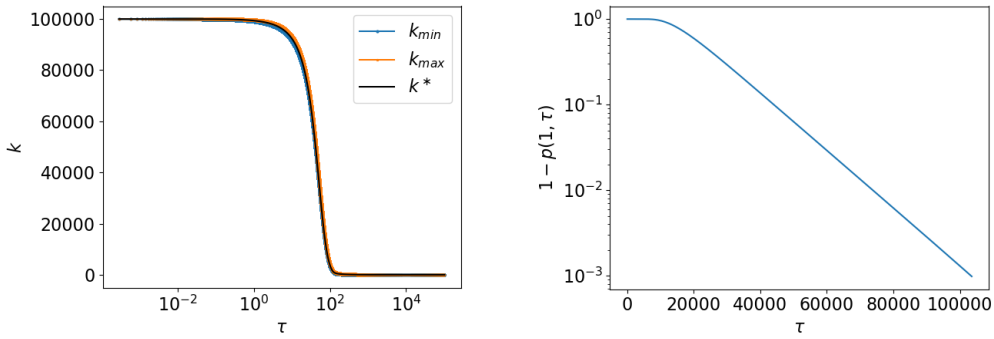


Figure 4.1: The ancestral lens with $n = 10^5$ and demographic model 2. The second plot shows the probability of (non)coalescence as a function of ancestral time τ.

binary merger in the interval $[\tau, \tau + d\tau]$ is

$$p(k,\tau)\frac{k(k-1)}{2N(\tau)}d\tau + \mathcal{O}\left(d\tau^2\right).$$

Likewise, the probability that the ancestral sample is of size $k+1$ and undergoes a binary merger in $[\tau, \tau + d\tau]$ is

$$p(k+1,\tau)\,\frac{(k+1)k}{2N(\tau)}\,d\tau + \mathcal{O}\left(d\tau^2\right).$$

Therefore,

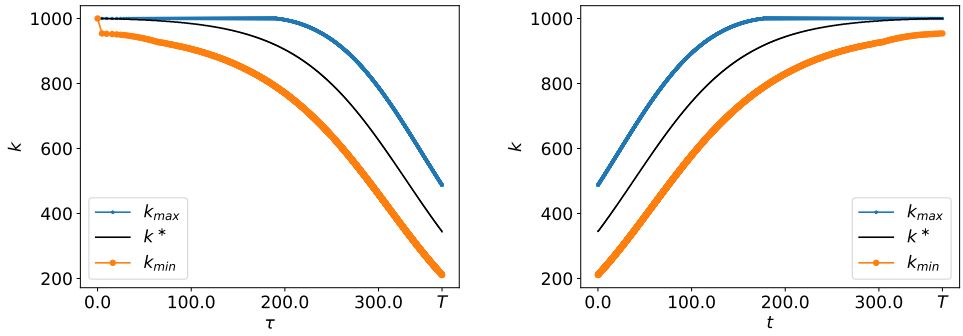
$$p(k,\tau+d\tau) - p(k,\tau) = -p(k,\tau)\,\frac{k(k-1)}{2N(\tau)}\,d\tau + p(k+1,\tau)\,\frac{(k+1)k}{2N(\tau)}\,d\tau + \mathcal{O}\left(d\tau^2\right).$$

In the limit $d\tau \to 0$, we obtain the differential equation

$$\frac{dp(k,\tau)}{d\tau} = -\frac{k(k-1)}{2N(\tau)}p(k,\tau) + \frac{k(k+1)}{2N(\tau)}p(k+1,\tau). \tag{4.5}$$

To calculate the ancestral lens, this differential equation is initialized with $p(n,0) = 1$ and $p(k,0) = 0$ for $k \neq n$. The numerical method used for computing $k_{min}(\tau)$ and $k_{max}(\tau)$ such that $p(k,\tau) < \epsilon_{lens}$ for $k \notin [k_{min}(\tau), k_{max}(\tau)]$ is described in Section 4.5. The functions $k_{min}(\tau)$ and $k_{max}(\tau)$ are the boundaries of the ancestral lens.

Figure 4.1 shows the ancestral lens for a demographic model. The considerable savings realized by confining calculations to the ancestral lens are obvious from that figure. We work with three demographic models:



(a) Ancestral lens

(b) Calendar lens

Figure 4.2: The ancestral and calendar lenses for $n = 1000$ samples under demographic model 1.

- Demographic model 0: N is assumed to be constant and equal to 2×10^4.
- Demographic model 1: $N(\tau) = N_0 e^{-r\tau}$, with $N_0 = 8 \times 10^6$ and $r = 0.017$, for $\tau \in [0, 367.8]$ and $N(\tau) = N_0 e^{-r367.8} \approx 15403$ for $\tau \geq 367.8$. This model is from [68].
- Demographic model 2: $N(\tau) = N_0 e^{-r\tau}$, with $N_0 = 8 \times 10^6$ and $r = 0.0538$, for $\tau \in [0, 119.47]$ and $N(\tau) = N_0 e^{-r119.47} \approx 12932$ for $\tau \geq 119.47$. This model is from [33].

Figure 4.1 uses model 2. In that figure, the stopping time $\tau = T$ for the ancestral lens is determined using the criterion $1 - p(1, T) \geq 10^{-3}$, which ensures that the entire sample has coalesced with probability greater than 99.9%. Such a criterion makes T very large, however, as shown in the figure. It is better to stop the lens when μ and N have become constant. The entire section of the lens at the stopping time T can be initialized using exact formulas for q_0, q_1, and the SFS.

The computation of the ancestral lens is forward in ancestral time τ. However, the computation of q_0, q_1, and the SFS march forward in calendar time. Therefore, the ancestral lens must be flipped to a calendar time lens as shown in Figure 4.2. In both Figures 4.1 and 4.2, k^* is the most likely size of the ancestral sample. We take

the stopping time $\tau = T$ to be the origin $t = 0$ of calendar time so that the current epoch is $t = T$.

4.2.2 Computation of $q_0(k,t)$ and $q_1(k,t)$

We denote the probability k samples at calendar time t, $t \in [0,T]$ coalesce without being hit by a mutation as $q_0(k,t)$. Additionally, we regard any exit from the ancestral lens as equivalent to coalescence. The probability of such an event is very low ($< \epsilon_{lens}$), so any loss of accuracy will be negligible compared with quadrature error. As a result, $q_0(k,t) = 1$ for $k \notin [k_{min}(t), k_{max}(t)]$. Also, $q_0(1,t) = 1$.

Suppose there are k ancestral samples at time t. If we go back to time $t - dt$ and ignore dt^2 terms, we have the following possibilities:

- The sample is hit with a mutation with probability $\mu(t)k\,dt$.
- There is a binary merger with probability $k(k-1)\,dt/2N(t)$.
- There is neither a mutation nor a binary merger with probability $1 - \mu(t)k\,dt - k(k-1)\,dt/2N(t)$.

Because we are ignoring dt^2 terms, these three possibilities are disjoint and exhaustive.

We have

$$q_0(k,t) = \left(1 - \mu(t)k\,dt - \frac{k(k-1)\,dt}{2N(t)}\right) q_0(k, t-dt) + \frac{k(k-1)}{2N(t)} q_0(k-1, t-dt).$$

If the k samples are neither hit by a mutation nor experience a binary merger in $[t-dt, t]$, we require that k samples at time $t-dt$ coalesce (or exit the lens) without a mutation. If the samples experience a binary merger, we require that $k-1$ samples at time $t-dt$ coalesce (or exit the lens) without a mutation. In the $dt \to 0$ limit, we

have the following differential equation:

$$\frac{dq_0(k,t)}{dt} = -\left(\mu(t)k + \frac{k(k-1)}{2N(t)}\right) q_0(k,t) + \frac{k(k-1)}{2N(t)} q_0(k-1,t). \tag{4.6}$$

The differential equation is solved for $t \in [0,T]$ as an initial value problem. For $k \in [k_{min}(0), k_{max}(0)]$, $q_0(k,0)$ is initialized using the exact formula for $q_0(k)$ when μ and N are constant. After solving the initial value problem, we obtain $q_0(n,T)$, which is the probability that n samples at the current epoch coalesce (or exit the lens) without being hit by a mutation.

Similarly, let $q_1(k,t)$ be the probability that k ancestral samples at time t coalesce (or exit the lens) while being hit by exactly one mutation. In this case, the same three disjoint and exhaustive possibilities imply

$$q_1(k,t) = \left(1 - \mu(t)k\,dt - \frac{k(k-1)\,dt}{2N(t)}\right) q_1(k,t-dt) + \mu(t)k\,dt\,q_0(k,t-dt) + \frac{k(k-1)}{2N(t)}\,q_1(k-1,t).$$

If there is neither a binary merger nor a mutation in $[t-dt, t]$, we require that k samples at time $t-dt$ coalesce (or exit the lens) with exactly one mutation. If there is a mutation event, we require that k samples at time $t-dt$ coalesce (or exit the lens) without suffering a mutation. Finally, if there is binary merger, we require that $k-1$ samples at time t coalesce (or exit the lens) while suffering exactly one mutation.

In the limit $dt \to 0$, we have the differential equation

$$\frac{dq_1(k,t)}{dt} = -\left(\mu(t)k + \frac{k(k-1)}{2N(t)}\right) q_1(k,t) + \mu(t)k q_0(k,t) + \frac{k(k-1)}{2N(t)} q_1(k-1,t). \tag{4.7}$$

This differential equation too is solved as an initial value problem from $t = 0$ to $t = T$. For $k \in [k_{min}(0), k_{max}(0)]$, $q_1(k,0)$ is initialized using the exact formula for $q_1(k)$ with constant μ and N. If $k \notin [k_{min}(t), k_{max}(t)]$, then $q_1(k,t) = 0$. In addition, $q_1(1,t) = 0$ for $t \in [0,T]$.

4.2.3 Computation of the SFS under the condition $\mathcal{C}_n$

Suppose the probability that j samples out of n are mutants at a polymorphic site is p_j. We may then represent the SFS using a generating function as $\sum_{j=1}^{n-1} p_j x^j$. The generating function representation will be used to define an extension operator and to derive a differential equation for the SFS.

Suppose the generating function is denoted S and that $n+1$ samples are related to the n samples via a binary merger. To end up with j mutants among $n+1$ samples, we must have either j mutants among n samples with one of the non-mutants splitting (which occurs with probability $(1 - j/n)$) or $j - 1$ mutants among n samples with one of the mutants splitting (which occurs with probability $(j - 1)/n$). Therefore, the probability, p'_j, that j out of $n + 1$ samples are mutants is given by

$$p'_j = p_j \left(1 - \frac{j}{n}\right) + p_{j-1}\frac{j-1}{n}$$

for $j = 2, \ldots, n-1$, by $p'_1 = p_1(1-1/n)$ for $j = 1$, and $p'_n = p_{n-1}(n-1)/n$ for $j = n$. We define the extension operator $\mathcal{E}$ in the following way

$$\mathcal{E}S = \sum_{j=1}^{n} p'_j x^j.$$

The extension operator $\mathcal{E}$ applied to the SFS S results in the SFS of a sample of size greater by one that is assumed to be related to the parental sample via a single binary merger.

Suppose $k \in [k_{min}(t), k_{max}(t)]$. Under the condition $\mathcal{C}_k$ that those k ancestral samples coalesce (or exit the lens) with exactly one mutation, the k samples are hit with a mutation in the time interval $[t - dt, t]$ with a probability equal to

$$\frac{k\mu(t)\, dt\, q_0(k,t)}{q_1(k,t)}.$$

Thus, under the condition $\mathcal{C}_k$, the k samples are hit with a mutation at the rate

$$p(\mu|k,t) = \frac{k\mu(t)q_0(k,t)}{q_1(k,t)}. \tag{4.8}$$

Similarly, the k samples experience a binary merger with the conditional rate given by

$$p(\beta|k,t) = \frac{k(k-1)}{2N(t)}\,\frac{q_1(k-1,t)}{q_1(k,t)}. \tag{4.9}$$

It must be noted that $p(\mu|k,t)$ and $p(\beta|k,t)$ are rates and not probabilities.

Under the condition $\mathcal{C}_k$, there are three disjoint and exhaustive possibilities (with dt^2 terms ignored) for k ancestral samples in the time interval $[t-dt, t]$.

- The k ancestral samples are hit with a mutation with probability $p(\mu|k,t)\,dt$.
- The samples experience a binary merger with probability $p(\beta|k,t)\,dt$.
- There is no event with probability $(1 - p(\mu|k,t) - p(\beta|k,t))\,dt$.

Let $S(k,t)$, with $t \in [0,T]$ and $k \in [k_{min}(t), k_{max}(t)]$, denote the SFS of k ancestral samples at time t under the condition $\mathcal{C}_k$. The three disjoint possibilities listed above imply that

$$S(k,t) = (1-p(\mu|k,t)-p(\beta|k,t))\,dt\,S(k,t-dt)+p(\mu|k,t)\,dt\,x+p(\beta|k,t)\,dt\,\mathcal{E}S(k-1,t-dt).$$

The middle term corresponds to a single mutant arising because of a mutation occurring in the interval $[t-dt, t]$. The last term corresponds to a binary merger.

In the limit $dt \to 0$, we obtain the differential equation

$$\frac{dS(k,t)}{dt} = -(p(\mu|k,t) + p(\beta|k,t))S(k,t) + p(\mu|k,t)x + p(\beta|k,t)\mathcal{E}S(k-1,t). \tag{4.10}$$

This differential equation is solved as an initial value problem from $t = 0$ to $t = T$. The SFS $S(k,0)$ for $k \in [k_{min}(0), k_{max}(0)]$ is initialized using the exact formula when

μ and N are both constant. For $k \notin [k_{min}(t), k_{max}(t)]$, $S(k,t) = 0$. In addition, $S(1,t) = 0$. The numerical solution of this differential equation is described in Section 4.5. A computer program implementing the algorithm may be obtained from `github.com/divakarvi/18-varymu`.

Suppose $\mu(t) = \epsilon\nu(t)$ and we take $\epsilon \to 0$. The resulting limit is the zero mutation limit with varying mutation rate. An algorithm applicable to this limit is described in Section 4.5.

4.2.4 The effect of non-binary mergers on the SFS

The coalescent assumes every merger to be a binary merger. That assumption only holds if $n \ll N^{1/3}$. If the sample size n is large, the same parent may have multiple children and there may be multiple parents with two or more children from among the samples over a single generation. [33] considered whether the SFS computed assuming every merger to be a binary merger is reliable. In addition, they derive the merger rate of $n(n-1)/2N$ assumed by the coalescent for $n \ll N^{1/2}$.

The SFS under the coalescent may be compared to the SFS under the Wright-Fisher model to understand if the two assumptions noted above have an effect on the SFS for large n. The formal truncation error in passing from the Wright-Fisher model to the coalescent is n^2/N [11]. However, the leading truncation term in the SFS is only n/N and in fact the total variation distance in the SFS is only around 1% for $n = N = 20{,}000$ as seen in Chapter 3. Furthermore, in demographic models such as 1 and 2 characterized by recent exponential growth, the total variation distance may perhaps be expected to be even lower. Assuming every merger to be a binary merger appears to have little effect on the SFS under the assumption of a single mutation in the genealogy. How the SFS under Wright-Fisher and the coalescent differ if the sample size n is large enough to make multiple mutations in the genealogy likely is

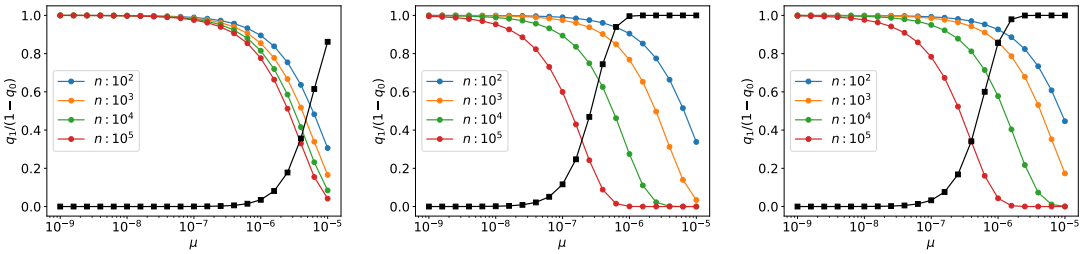


Figure 4.3: Probability that a site is polymorphic as a result of a single mutation in the genealogy for demographic models 0, 1, 2 and various sample sizes. In all three plots, the black squares plot the probability for $n = 10^5$ samples that a site hit with a mutation has been hit with three or more mutations.

not known.

In Section 4.5, we show how to compute the SFS with varying $\mu(t)$ and $N(t)$ under the Wright-Fisher model. An advantage of the Wright-Fisher model is that it is already discrete. We also remark that computations with the Wright-Fisher model may lend themselves to better optimization with the use of suitable asymptotic formulas.

4.3 Visualization and Results

4.3.1 The effects of constant nonzero mutation rate on the SFS

[33] have presented evidence that the infinite sites model is violated for samples of around $n = 10^5$ haploid human genomes. The infinite sites models assumes that every new mutation in the genealogy occurs at a different site.

To understand when and how violations of the infinite sites model set in, we may first look at the quantity

$$\frac{q_1(n,0)}{1-q_0(n,0)},$$

which is the probability that there is exactly one mutation in the genealogy given that there are one or more mutations in the genealogy of the n samples. In Figure 4.3, we use $q_1/(1-q_0)$ as a surrogate for the probability that a site is polymorphic as

a result of a single mutation. The two probabilities will be close but are not exactly the same. They differ slightly because of the small probability that two mutations may occur in the same lineage in the genealogy and cancel each other. Thus, a sample with two or mutations in its genealogy may not be polymorphic.

In Figure 4.3, we have graphed the probability of a single mutation at a polymorphic site for demographic model 0 which assumes $N \equiv 2 \times 10^4$. Sample sizes of $n = 10^5$ would not make sense in the Wright-Fisher interpretation of that model. However, in the coalescent N is only a parameter to control rates of binary mergers.

The probability of a single mutation at a polymorphic site is the highest in demographic model 0. That appears to be because binary mergers are initially fastest in demographic model 0.

Both demographic models 1 and 2 assume exponentials that persist for more than a 100 generations. The population explosion slows down binary mergers and as a result the probability of double mutations is higher in demographic models 1 and 2. The exponential persists over a greater interval of time in model 1 and model 1 shows a greater probability of double mutations than model 2. In both models 1 and 2, the probability of a multiple mutation is noticeable for even $\mu = 10^{-8}$ and $n = 10^5$.

The probability that a site hit with a mutation has been hit with three or more mutations is given by

$$\frac{1 - q_0(n,0) - q_1(n,0) - q_2(n,0)}{1 - q_0(n,0)}.$$

For demographic model 1, $n = 10^5$, and $\mu = 10^{-7}$, the probability that a site that has been hit with a mutation has been hit with exactly one, exactly two, or three or more mutations is 60%, 28.4%, and 11.6%, respectively (see Figure 4.3). For demographic model 2, those probabilities are 78.3%, 18.4%, and 3.3%, respectively.

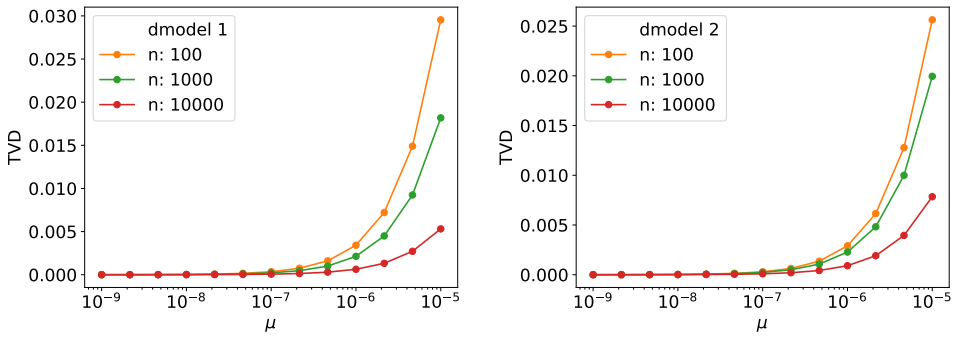


Figure 4.4: The total variation distance between the SFS at a given μ and at $\mu = 0$ under the condition $\mathcal{C}_n$.

[33] have inferred a mutation rate of around 10^{-7} for CpG junctions. The SFS will be a mixture of the SFS due to a single mutation, the SFS due to two mutations, and the SFS due to three or more mutations in the genealogy. A sample being hit with three mutations at a site does not necessarily imply that the site shows four-fold polymorphism. Because transitions are more likely than transversions, the site may only be biallelic.

Suppose a sample carries multiple mutations at the same site in its genealogy. The mutations may not be nested in the genealogical tree, in general. However, for large n mutations being nested is not a likely scenario.

Assuming mutations are not nested, we may disentangle the effects of single and multiple mutations on the SFS. To do so, we turn to Figure 4.4. The total variation distance (or variation distance) between an SFS given by p_j and an SFS given by p'_j for $j = 1, \ldots, n-1$ is defined as

$$\frac{1}{2} \sum_{j=1}^{n-1} \left| p_j - p'_j \right| .$$

It is the right metric to use for comparison because it can be interpreted as the maximum difference between the probabilities of any possible event that is a subset of $\{1, \ldots, n-1\}$ [6]. Because it can be interpreted as a probability, the variation

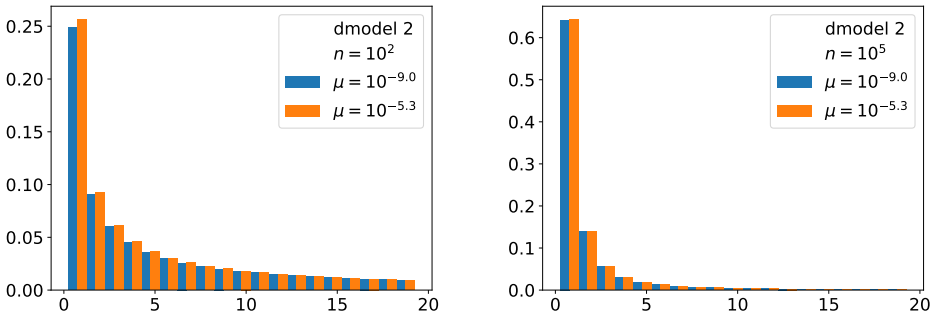


Figure 4.5: The effect of finite μ on the SFS ($j < 20$).

distance can also be thought of as a percent.

Figure 4.4 shows that the variation distance of the SFS with μ finite and $\mu = 0$ under the condition $\mathcal{C}_n$ falls as the sample size n increases. The variation distance is negligible for $\mu = 10^{-7}$. Thus, the phenomena described by [33] are entirely due to multiple mutations in the genealogy.

In addition to the variation distance of the SFS between μ finite and $\mu = 0$ being small, it is in a direction opposite to the total effect. Figure 4.5 shows that the effect of finite μ is to slightly increase the occurrence of rare alleles. However, in the overall SFS, rare alleles are depleted [33].

Intuitively, we may understand why rare alleles are depleted by multiple mutations as follows. The probability of singletons ($j = 1$ mutants) begins to dominate in large samples as evident from Figure 4.5. When the genealogy carries two or more mutations, singletons cannot occur when there are multiple non-nested mutations in the genealogy. Thus, the probability of singletons is depleted by an amount approximately equal to $q_1/1 - q_0$ or the probability of multiple mutations in the genealogy of a sample.

4.3.2 The effects of varying mutation rate on the SFS

For purposes of analyzing the effects of variable mutation rate, we use as a simple test model the mutation rate given by

$$\mu(\tau) = \begin{cases} \mu_0 \frac{T-\tau}{T} + \frac{\mu_0}{f}\frac{\tau}{T}, & \tau \in [0, T] \\ \mu_0/f & \tau \geq T \end{cases} \tag{4.11}$$

The mutation rate is assumed to vary linearly in $[0, T]$. In addition, we assume $T = 367.8$ and $T = 119.47$ for demographic models 1 and 2, respectively. Following [79], we take $\mu_0 = 1.2 \times 10^{-8}$.

In the model for $\mu(t)$, the variation in μ sets in at $\tau = T$. For demographic models 1 and 2, we have assumed T to be to the epoch at which exponential increase in population sets in. In the model, f is the factor by which the mutation rate increases from $\tau = T$, which is T generations in the past, to $\tau = 0$, which is the current time.

Methods used to infer the mutation rate rely directly or indirectly on the pioneering work of [94] on the number of segregating sites. For a recent example, see [**?**]. Therefore, it is appropriate to begin by looking at the probability $1 - q_0(n, 0)$ that a current sample of size n has been hit with a mutation at a site. The probability $1 - q_0(n, 0)$ is close to but not exactly the same as the probability that a site is segregating. If a site is hit with multiple mutations, there is a small probability that it is not segregating.

Figure 4.6 shows the way the number of segregating sites (more precisely, the probability a site is hit with a mutation) varies as a function of f. A sample size of $n = 100$ appears more sensitive to variations in the mutation rate than larger samples, especially when $f > 1$ and $\mu(t)$ is increasing.

Figure 4.7 shows the total variation distance of the SFS between $f = f$ and $f = 0$

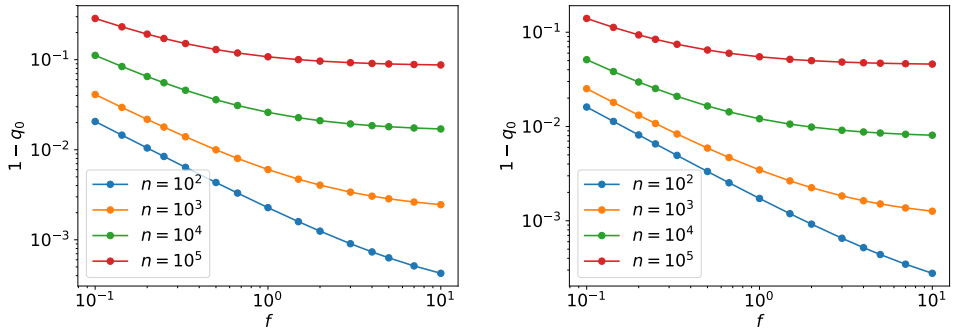


Figure 4.6: The probability that a site is segregating (more precisely, hit with a mutation) as a function of the factor f (see (4.11)) for demographic models 1 and 2, respectively.

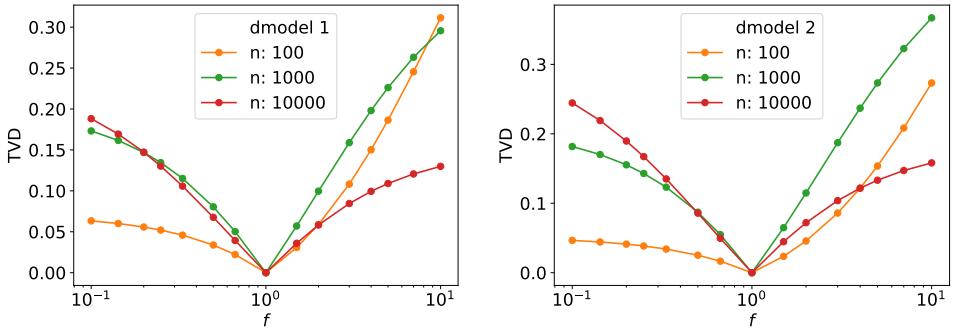


Figure 4.7: Total variation distance of the SFS for a given f (see (4.11)) from the SFS with $f = 0$, which implies a constant mutation rate. The parameter f controls the variation in mutation rate.

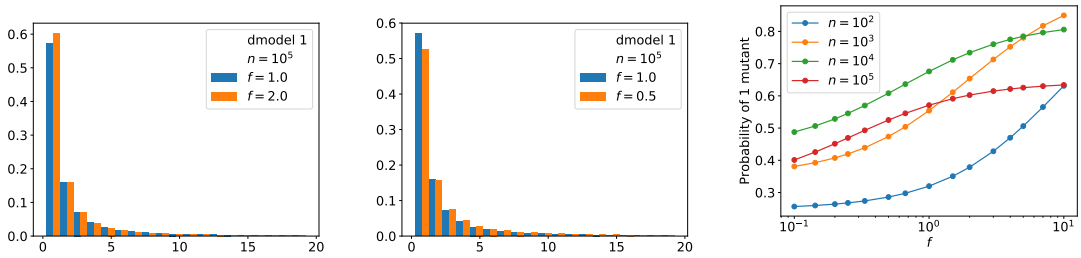


Figure 4.8: The effect of increasing and decreasing mutation rates on the SFS. The last plot uses demographic model 1 as well.

as a function of the parameter f. There is no easily discernible pattern in the plots. For the demographic model 1 and $n = 100$, the SFS changes by 6% and 3% when the mutation rate is assumed to double or halve, respectively. For demographic model 2, those numbers are 5% and 3%, respectively. The variation in $\mu(t)$ certainly has an effect on the SFS. Figure 4.8 shows that an increasing mutation rate augments the fraction of rare alleles, whereas a decreasing mutation rate depletes it.

The effect of varying mutation rate on the SFS is not necessarily the greatest for the largest samples. The third plot in Figure 4.8 shows that the effect is greater for $n = 10^3$ than $n = 10^5$ under demographic model 2.

The effect of varying mutation rate on the SFS may not be as great as the effect of varying demography. Yet, the fact that differences in mutation rates between human populations can be detected is evidence that the variation of the mutation rate over human history can be detected. Our method for computing the SFS may prove useful in that regard.

4.4 Discussion

The Polanski-Kimmel algorithm for computing the SFS [74] allows the population size $N(t)$ to vary but assumes the mutation rate to be constant and negligible. We have presented an algorithm to compute the SFS that allows the mutation rate $\mu(t)$ to

be finite and varying. The main innovation in our algorithm is to ignore the internal structure of the genealogical tree and instead take a more Markovian approach.

The algorithm uses a first pass (with ancestral time τ increasing) to calculate the ancestral lens. The ancestral lens is then flipped to the calendar time lens. In the second pass, the SFS as well as $q_0(k,t)$ and $q_1(k,t)$ are calculated with calendar time t increasing. The calculations solve a set of linear ordinary differential equations with time-varying coefficients.

The quantities $q_0(k,t)$ and $q_1(k,t)$ are the probabilities that a sample of size k coalesces without being hit by a mutation or after being hit by exactly one mutation, respectively. The algorithms for calculating q_0 and q_1 may be thought of as generalizations of the work of [94] on the number of segregating sites. The generalization consists in allowing both $N(t)$ and $\mu(t)$ to vary arbitrarily.

When the sample size is $n \approx 10^5$ and the mutation rate is $\mu \approx 10^{-7}$, multiple mutations occur in human genealogies with appreciable probabilities. There could even be three mutations in the genealogy. Algorithms for obtaining the SFS with two mutations have been applied to sample sizes of $n = 100$ [41]. In addition to reaching greater sample sizes, possibly while limiting calculations to rare alleles, there are yet other issues to be investigated. One such issue is the SFS under the assumption of three mutations in the genealogy.

4.5 Implementation Details

In this section, we explain how the ancestral lens and the SFS are computed numerically. We also give a version of the algorithm that allows $\mu(t)$ to vary and then takes the zero mutation limit. In addition, we show how to calculate the SFS for the Wright-Fisher model with varying mutation rate and population size.

4.5.1 Implementation of the ancestral lens

The differential equation (4.5) is linear with non-constant coefficients. The decay rate $k(k-1)/2N(\tau)$ is around 10^4 for $k \approx 10^5$ and $N \approx 10^6$. Yet, the differential equations for $k = 2, \ldots, n$ cannot be considered stiff because the decay rate corresponds to the rate of binary mergers and must be resolved. Out of precaution, we used a 4th order BDF discretization [31]. There is no additional cost to using an implicit method because the equations are linear. Initial time-steps were taken using the implicit midpoint rule. The formal order of accuracy is 3.

Integrating the differential equations for $k = 2, \ldots, n$ would be too expensive for large n. Instead, we restrict the integration to the ancestral lens even as it is being computed. Suppose the ancestral lens at τ is given by $[k_{min}, k_{max}]$. In the time step from τ to $\tau+h$, we use $k = \max(k_{min}-1, 1), \ldots, k_{max}$. After the time step, $k_{min}(\tau+h)$ can be either k_{min} or $k_{min} - 1$. It is equal to $k_{min} - 1$ if $p(k_{min} - 1, \tau + h) < 10^{-40}$ and k_{min} otherwise. Similarly, $k_{max}(\tau + h)$ can be either $k_{max} - 1$ or k_{max}. It is equal to $k_{max} - 1$ if $p(k_{max}, \tau + h) < 10^{-40}$ and k_{max} otherwise.

Particular care is necessary during the very first time step. The ancestral lens at $\tau = 0$ is given by $k \in [n, n]$ and consists of a single point. If the above strategy is followed, the ancestral lens can grow to at most $k \in [n-1, n]$ after the first time step. The tolerance of $\epsilon_{lens} = 10^{-40}$ is so small that the ancestral lens will in fact be much wider. Failing to capture its width correctly in the first step will corrupt the entire computation. To capture the width of the ancestral lens correctly, the implicit midpoint rule is iterated until the width of the ancestral lens stabilizes. The way the ancestral lens grows during the very first step may be observed from Figure 4.2a.

Figure 4.2b shows an ancestral lens flipped to a calendar time lens. The flip is mostly straightforward except at the current epoch $t = T$ or $\tau = 0$. At $\tau = 0$,

the ancestral lens is given by $k \in [n, n]$. However, $k \in [n, n]$ at $t = T$ will not do because the differential equations (4.6), (4.7), and (4.10) utilize information regarding $k - 1$ samples. When an implicit BDF discretization is involved, narrowing the lens suddenly at $t = T$ will create error in moving information from $n - 1$ samples to n. This problem is easily solved by taking the calendar time lens at t to be equal to the ancestral lens at $\tau = T - t + h$, where h is the time step into t.

4.5.2 Solution of the differential equations for q_0, q_1, and the SFS

Provided the calendar time lens is calculated carefully, no really new issues arise in the solution of the differential equations (4.6), (4.7), and (4.10) for $q_0(k, t)$, $q_1(k, t)$, and $S(k, t)$, respectively. The differential equations are discretized using 4th order BDF with initial time steps using the implicit midpoint rule. The differential equations for $q_0(k, t)$, $q_1(k, t)$, and $S(k, t)$ are solved simultaneously. Therefore, the memory requirements of this algorithm are very low.

The differential equations are solved only for $k \in [k_{min}(t), k_{max}(t)]$. During the time step from t to $t + h$, k_{min} or k_{max} (or both) may increase by 1. If q_0 is assumed to be 1 and q_1, S are assumed to be zero outside the lens, as stated in the main text, no special handling is necessary when k_{max} increases. When k_{min} increases by 1, the values of $q_0(k_{min}, t)$ as well as $q_0(k_{min}, \cdot)$ at previous epochs that contribute to the step to $t + h$ are taken to be 1. Similarly, the corresponding values of q_1 and S are taken to be 0.

4.5.3 Choice of time step and accuracy

Suppose $dx/dt = -\alpha(t)x$. Then

$$\frac{d^4x}{dt^4} = (\alpha(t)^4 - 6\alpha(t)^2\dot{\alpha}(t) + 3\dot{\alpha}(t)^2 + 4\alpha(t)\ddot{\alpha}(t) - \dddot{\alpha}(t))x.$$

Because the numerical discretizations are formally of order 3, the time step h is obtained from the requirement

$$\left|\alpha(t)^4 - 6\alpha(t)^2\dot{\alpha}(t) + 3\dot{\alpha}(t)^2 + 4\alpha(t)\ddot{\alpha}(t) - \dddot{\alpha}(t)\right| h^4 \leq htol.$$

In this requirement, we have taken $x = 1$ because all the differential equations are solving for probabilities. To preserve the numerical stability of the BDF formula, each new time step must be within a factor of 1.2 of the previous time step.

In the computation of the ancestral lens, we take

$$\alpha(\tau) = -\frac{k(k-1)}{2N(\tau)}$$

with $k = k_{max}(\tau)$. In the computation of $q_0(k,t)$, $q_1(k,t)$, and $S(k,t)$, we take

$$\alpha(t) = \frac{k(k-1)}{2N(t)} + k\mu(t)$$

with $k = k_{max}(t)$. Derivatives of α are computed using differences.

The accuracy of our program has been checked by comparing with implementations of Polanski-Kimmel algorithm [5, **?**] and against coalescent simulations [38]. In addition, we wrote an independent program in Python that solves the differential equations for $k = 1, \ldots, n$ without limiting itself to an ancestral lens. The accuracy of the C program has been checked against the Python program. The Python program used `odeint()`, which is defined in the `scipy` library. All the reported computations have at least 4 digits of accuracy and often more than 10 digits of accuracy.

4.5.4 Zero limit with varying mutation rate

Our algorithm to compute the SFS simplifies slightly if we take $\mu(t) = \epsilon\nu(t)$ and then take $\epsilon \to 0$. The probability $q_0(k,t)$ that k ancestral samples at time t coalesce without being hit by a mutation is then $q_0(k,t) = 1 + \mathcal{O}(\epsilon)$. In the limit $\epsilon \to 0$, we have $q_0(k,t) \equiv 1$.

If we replace $q_1(k,t)$ by $\epsilon q_1(k,t)$ in (4.7) and take $\epsilon \to 0$, we obtain

$$\frac{dq_1(k,t)}{dt} = -\frac{k(k-1)}{2N(t)}q_1(k,t) + \nu(t)k + \frac{k(k-1)}{2N(t)}q_1(k-1,t).$$

Equations (4.8), (4.9), and (4.10) may still be used to compute the SFS, the only change being the replacement of $\mu(t)$ in (4.8) by $\nu(t)$.

4.5.5 The Wright-Fisher SFS with varying mutation rate and population size

Suppose the population size in the Wright-Fisher model is N_t at time t and suppose the mutation rate to be μ_t when the t-th generation begets the $t+1$st generation. The probability that k samples at time $t+1$ have j parents is given by

$$p_{k,j,N_t} = \left\{ {k \atop j} \right\} \left(1 - \frac{1}{N}\right) \cdots \left(1 - \frac{j-1}{N}\right) \frac{1}{N^{k-j}}.$$

Here $\left\{ {k \atop j} \right\}$ is a Stirling number of the second kind.

Suppose $q_0(k,t)$ is the probability that k samples in generation t coalesce without being hit by a mutation. Then

$$q_0(k,t+1) = (1-\mu_t)^k \sum_{j=1}^{k} p_{k,j,N_t} q_0(j,t).$$

Suppose $q_1(k,t)$ is the probability that k samples in generation t coalesce with exactly one mutation in their Wright-Fisher genealogy. Then

$$q_1(k,t+1) = k\mu_t(1-\mu_t)^{k-1} \sum_{j=1}^{k} p_{k,j,N_t} q_0(j,t) + (1-\mu_t)^k \sum_{j=1}^{k} p_{k,j,N_t} q_1(j,t).$$

Suppose $\mathcal{C}_k$ is the condition or event that k samples coalesce with exactly one mutation in their Wright-Fisher genealogy. Under the condition $\mathcal{C}_k$, the probability that a sample of size k is hit with a mutation during generation $t+1$ is

$$p\left(\mu \middle| k, t+1\right) = \frac{k\mu_t(1-\mu_t)^{k-1} \sum_{j=1}^{k} p_{k,j,N_t} q_0(j,t)}{q_1(k,t+1)}.$$

Similarly, the conditional probability that it has j parents in generation t but does not experience a mutation during generation $t+1$ is

$$p\left(j\middle|k,t+1\right)=\frac{(1-\mu_t)^k p_{k,j,N_t} q_1(j,t)}{q_1(k,t+1)}.$$

The recurrence for the SFS is then given by

$$\mathcal{S}\left(k,t+1\right)=p\left(\mu\middle|k,t+1\right)x+\sum_{j=1}^{k}p\left(j\middle|k,t+1\right)\mathcal{E}_k\mathcal{S}(j,t).$$

Here $\mathcal{E}_k S(j,t)$ is an extension of the SFS for j samples to k samples, assuming the k samples to be children of the j samples. A formula for $\mathcal{E}_k$ follows from the Wright-Fisher propagators derived in Chapter 3.

The Stirling numbers are the main source of difficulty in implementing the Wright-Fisher model. However, uniform asymptotic formulas [85] can be used to make the Wright-Fisher implementation much more efficient. The coalescent is not much more than an asymptotic approximation to Stirling numbers of the second kind. If k children must be divided between j parents, under Wright-Fisher, the children are first partitioned into j sets in one of $\left\{{k \atop j}\right\}$ ways and each partition is assigned to a parent. The coalescent produces a partition of k children between j parents using binary mergers. If $k-j=1$, the two partition distributions are the same. The approximation is not a particularly good one in general, when $k-j\gg 1$, although it works quite well for the SFS as seen in Chapter 3.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[2] V. Aggarwala and B. Voight. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48(4):349–355, 2016.

[3] V. Aggarwala and B. F. Voight. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature genetics*, 47(3):349, 2015.

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

[5] 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.

[6] P. Bremaud. *Markov Chains*. Springer, 1999.

[7] J. Carlson, A. E. Locke, M. Flickinger, M. Zawistowski, S. Levy, R. M. Myers, M. Boehnke, H. M. Kang, L. J. Scott, J. Z. Li, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nature communications*, 9(1):3753, 2018.

[8] 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.

[9] 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.

[10] 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.

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

[12] M. Ehrlich and R.-H. Wang. 5-methylcytosine in eukaryotic DNA. *Science*, 212(4501):1350–1357, 1981.

[13] B. Eldon and J. Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633, 2006.

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

[15] L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. Sousa, and M. Foll. Robust demographic inference from genomic and SNP data. *PLoS Genetics*, 9(10), 2013.

[16] A. Eyre-Walker and Y. Eyre-Walker. How much of the variation in the mutation rate along the human genome can be explained? *G3: Genes, Genomes, Genetics*, 4(9):1667–1670, 2014.

[17] R. Fisher. On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341, 1922.

[18] R. Fisher. The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh*, 50:204–219, 1930.

[19] W. Fu, T. O'Connor, G. Jun, H. Kang, G. Abecasis, S. Leal, S. Gabriel, D. Altshuler, J. Shendure, D. Nickerson, M. Bamshad, and J. Akey. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, 2013.

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

[21] Y.-X. Fu. Statistical properties of segregating sites. *Theoretical Population Biology*, 48:172–197, 1995.

[22] Y.-X. Fu and W.-H. Li. Statistical tests of neutrality of mutations. *Genetics*, 133:693–709, 1993.

[23] Z. Gao, M. Wyman, G. Sella, and M. Przeworski. Interpreting the dependence of mutation rates on age and time. *PLoS Biology*, 14(1), 2016.

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

[25] 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.

[26] R. Griffiths. The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theoretical Population Biology*, 64(2):241–251, 2003.

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

[28] R. 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.

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

[30] R. Gutenkunst, R. Hernandez, S. Williamson, and C. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5, 2009.

[31] E. Hairer, S. Norsett, and G. Wanner. *Solving Ordinary Differential Equations I.* Springer-Verlag, Berlin, 1991.

[32] J. Haldane. The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Annals of Eugenics*, 13:262–271, 1946.

[33] A. Harpak, A. Bhaskar, and J. Pritchard. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS Genetics*, 12, 2016.

[34] K. Harris. Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences of the United States of America*, 112(11):3439–3444, 2015.

[35] G. Hellenthal, G. B. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, 2014.

[36] A. Hodgkinson and A. Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12:756–766, 2011.

[37] A. Hodgkinson, E. Ladoukakis, and A. Eyre-Walker. Cryptic variation in the human mutation rate. *PLoS Biology*, 7:0226–0232, 2009.

[38] R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, 2002.

[39] D. Hwang and P. Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39):13994–14001, 2004.

[40] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

[41] P. Jenkins, J. Mueller, and Y. Song. General triallelic frequency spectrum under demographic models with variable population size. *Genetics*, 196(1):295–311, 2014.

[42] P. Johnson and I. Hellmann. Mutation rate distribution inferred from coincident SNPs and coincident substitutions, 2011.

[43] J. Kamm, J. Terhorst, and Y. Song. Efficient computation of the joint sample frequency spectra for multiple populations. *Journal of Computational and Graphical Statistics*, 26:182–194, 2017.

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

[45] A. Keinan and A. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, 2012.

[46] 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.

[47] S.-H. Kim, N. Elango, C. Warden, E. Vigoda, and S. Yi. Heterogeneous genomic molecular clocks in primates. *PLoS Genetics*, 2(10):1527–1534, 2006.

[48] M. Kimura. Solution of a process of random genetic drift with a continous model. *Proc. N. A. S.*, 41:144–150, 1955.

[49] M. Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1:177–232, 1964.

[50] J. Kingman. Origins of the coalescent: 1974-1982. *Genetics*, 156:1461–1463, 2000.

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

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

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

[54] P.-R. Loh, M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich, and B. Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, pages genetics–112, 2013.

[55] S. Lukic and J. Hey. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-africa expansion. *Genetics*, 192:619–639, 2012.

[56] I. Mathieson and D. Reich. Differences in the rare variant spectrum among human populations. *PLoS Genetics*, 13(2), 2017.

[57] S. Matuszewski, M. Hildebrandt, G. Achaz, and J. Jensen. Coalescent processes with skewed offspring distributions and nonequilibrium demography. *Genetics*, 208:323–338, 2018.

[58] A. Melfi and D. Viswanath. Single and simultaneous binary mergers in Wright-Fisher genealogies. *Theoretical Population Biology*, 121:60– 71, 2018.

[59] A. Melfi and D. Viswanath. The site frequency spectrum under finite and time-varying mutation rates. *biorxiv.org*, 2018.

[60] A. Melfi and D. Viswanath. The Wright-Fisher site frequency spectrum as a perturbation of the coalescent's. *Theoretical Population Biology*, 2018.

[61] J. Michaelson, Y. Shi, M. Gujral, H. Zheng, D. Malhotra, X. Jin, M. Jian, G. Liu, D. Greer, A. Bhandari, W. Wu, R. Corominas, A. Peoples, A. Koren, A. Gore, S. Kang, G. Lin, J. Estabillo, T. Gadomski, B. Singh, K. Zhang, N. Akshoomoff, C. Corsello, S. McCarroll, L. Iakoucheva, Y. Li, J. Wang, and J. Sebat. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, 151(7):1431–1442, 2012.

[62] M. Möhle. Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models. *Advances in Applied Probability*, 32:983–993, 2000.

[63] M. Möhle and S. Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Annals of Probability*, 29:1547–1562, 2001.

[64] P. Moorjani, C. Amorim, P. Arndt, and M. Przeworski. Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences of the United States of America*, 113(38):10607–10612, 2016.

[65] P. Moorjani, Z. Gao, and M. Przeworski. Human germline mutation and the erratic evolutionary clock. *PLoS Biology*, 14(10), 2016.

[66] P. Moorjani, N. Patterson, J. N. Hirschhorn, A. Keinan, L. Hao, G. Atzmon, E. Burns, H. Ostrer, A. L. Price, and D. Reich. The history of african gene flow into southern europeans, levantines, and jews. *PLoS genetics*, 7(4):e1001373, 2011.

[67] V. Narasimhan, R. Rahbari, A. Scally, A. Wuster, D. Mason, Y. Xue, J. Wright, R. Trembath, E. Maher, D. Van Heel, A. Auton, M. Hurles, C. Tyler-Smith, and R. Durbin. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nature Communications*, 8(1), 2017.

[68] M. Nelson, D. Wegmann, M. Ehm, D. Kessner, P. St. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. Hall, K. Nangle, J. Wang, G. Abecasis, L. Cardon, S. Zöllner, J. Whittaker, S. Chissoe, J. Novembre, and V. Mooser. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 336(6090):100–104, 2012.

[69] J. P. Noonan, G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Pääbo, J. K. Pritchard, et al. Sequencing and analysis of neanderthal genomic dna. *science*, 314(5802):1113–1118, 2006.

[70] P. Palamara. ARGON: Fast, whole-genome simulation of the discrete time Wright-Fisher process. *Bioinformatics*, 32(19):3032–3034, 2016.

[71] P. Palamara, J. Terhorst, Y. Song, and A. Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50:1311–1317, 2018.

[72] N. J. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, pages genetics–112, 2012.

[73] A. Polanski, A. Bobrowski, and M. Kimmel. A note on distributions of times to coalescence, under time-dependent population size. *Theoretical Population Biology*, 63(1):33–40, 2003.

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

[75] 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.

[76] R. Rahbari, A. Wuster, S. Lindsay, R. Hardwick, L. Alexandrov, S. Al Turki, A. Dominiczak, A. Morris, D. Porteous, B. Smith, M. Stratton, and M. Hurles. Timing, rates and spectra of human germline mutation. *Nature Genetics*, 48(2):126–133, 2016.

[77] D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, et al. Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, 468(7327):1053, 2010.

[78] S. Sawyer and D. Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176, 1992.

[79] A. Scally and R. Durbin. Revising the human mutation rate: Implications for understanding human evolution. *Nature Reviews Genetics*, 13(10):745–753, 2012.

[80] K. Schwarze, J. Buchanan, J. C. Taylor, and S. Wordsworth. Are whole-exome and whole-genome sequencing approaches cost-effective? a systematic review of the literature. *Genetics in Medicine*, 2018.

[81] M. Steinrücken, E. M. Jewett, and Y. S. Song. SpectralTDF: transition densities of diffusion processes with time-varying selection parameters, mutation rates and effective population sizes. *Bioinformatics*, 32(5):795–797, 2016.

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

[83] F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123:585–595, 1989.

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

[85] N. Temme. Asymtotic estimates of Stirling numbers. *Studies in Appl. Math.*, 89:233–243, 1993.

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

[87] L. Trefethen. *Approximation Theory and Approximation Practice*. SIAM, 2013.

[88] D. Živković, M. Steinrücken, Y. Song, and W. Stephan. Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics*, 200(2):601–617, 2015.

[89] J. Wakeley. *Coalescent Theory*. W.H. Freeman, 2009.

[90] J. Wakeley and J. Hey. Estimating ancestral population parameters. *Genetics*, 145:847–855, 1997.

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

[92] 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.

[93] B. Waltoft and A. Hobolth. Non-parametric estimation of population size changes from the site frequency spectrum. *Statistical Applications in Genetics and Molecular Biology*, 2018.

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

[95] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.