RECONSTRUCTING POPULATION HISTORIES THROUGH THE CONDITIONAL GENEALOGY UNDER A MULTIPLE MERGER COALESCENT MODEL

ALICIA ZHANG

Abstract. An important goal in population genetics is to infer the history of populations using present-day genetic data. Here, we examine the effect of highly reproductive events (HREs) within a diploid population-genetic model, where occasionally a single pair of individuals has some number of offspring on the order of the population size. Specifically, we study the case where the population size tends to infinity. The gene genealogy of a sample of the population is characterized by a standard Kingman coalescent interrupted by HREs whose times are determined by a Poisson point process. First, we present an algorithm for simulating this model. Next, we examine specifically the expected height of the coalescent tree conditioned on the first few HRE time points, as opposed to the whole process. Building upon previous literature, we compute the expected pairwise coalescence time conditioned on the first few time points. By combining (1) our knowledge of the number of time points necessary to determine the expected height up to some small error and (2) our computation of this height in terms of these time points, we hope to recover HREs from real-life genomic data.

Contents

Date: September 9, 2024.

1. Introduction

1.1. **Motivation.** Population genetics is a branch of genetics that studies genetic variation within populations over time and space. One of the foundational models in this field is coalescent theory, which traces the genealogical history of alleles within a population. It provides a framework for understanding how all alleles at a particular locus in a population can be traced back to a single common ancestor, known as the most recent common ancestor (MRCA). By tracing the lineage of genes back through generations, coalescent theory helps us reconstruct the gene genealogy, offering insights into the evolutionary history of populations.

Coalescent theory is particularly useful in studying the effects of demographic events, such as population bottlenecks or periods of reduced genetic diversity, which can significantly impact the structure of a population's gene pool. An example of such a population bottleneck might be a "big family" event, where a small number of individuals contribute disproportionately to the gene pool of future generations. Such events, often influenced by random genetic drift or selection, can lead to reduced genetic variation and increased relatedness within a population, ultimately shaping its genetic landscape.

In this report, we will explore a multiple merger coalescent model that is characterized by big family events. Our focus will be on analyzing specific parameters of the model's coalescent tree, with particular attention to its height, which represents the time to the most recent common ancestor. We hope that the height is discernable from genomic data, so that we can answer the question: can we recover big family events in population history by examining genomic data?

1.2. **Model introduction.** We will be using a "bursts of coalescence model" as defined in (DFBW24). In the discrete case, this model is characterized by two different types of reproduction, where population size N is finite and generations are nonoverlapping. In each generation, with some small probability α_N, there is some highly reproductive (HR) couple whose offspring are a fraction ψ of the population in the next generation. Call each of these events HREs (Highly Reproductive Events). Otherwise, the classic diploid biparental Wright-Fisher model occurs with probability $1-\alpha_N$. Figure 1 and 2 each display an example of a Wright-Fisher and an HRE generation, respectively.

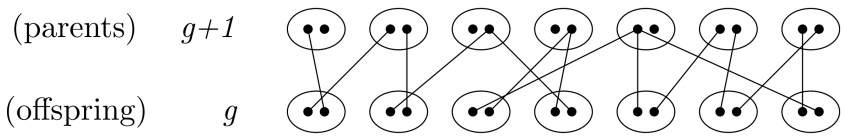


Figure 1. Wright-Fisher generation, $N = 7$

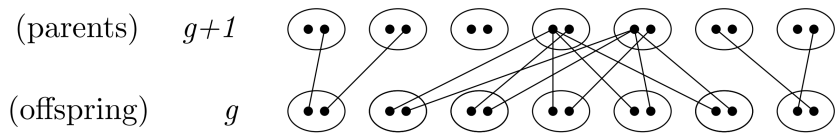


Figure 2. HRE generation, $N = 7, [\psi N] = 5$

2. Theory and Methods

2.1. **Limiting case as $N \to \infty$.** As the population size tends to infinity, we sample n loci from distinct individuals in generation $g = 0$ and trace down their lineage.

Here, the big family events occur according to a Poisson process $\vec{t} = t_1, t_2, \cdots, t_i, \cdots$ with some rate λ. In between these events, the standard Kingman coalescent occurs.

The Kingman coalescent is the limiting case of the Wright-Fisher model. Under a diploid population, coalescence behavior is characterized as such: each pair of lineages coalesces with rate $\frac{1}{2}$, independently of other pairs.

Figure 3 is an example of a coalescent tree of the limiting HRE model. HRE events are shown using dotted lines.

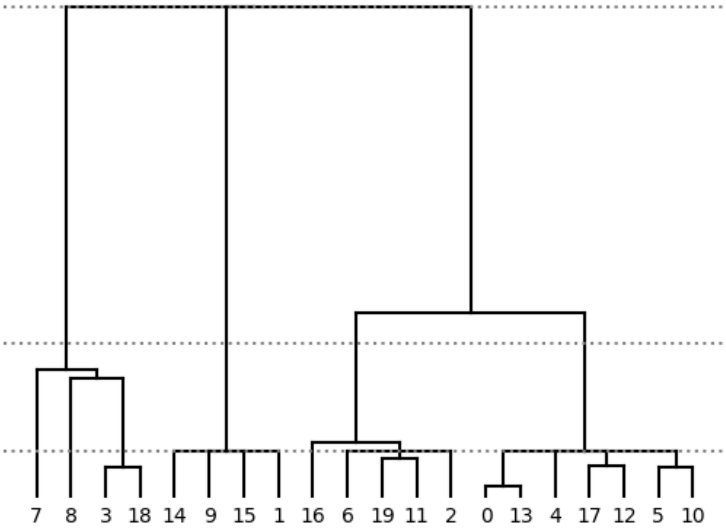


Figure 3. HRE model coalescent tree ($n = 20$, $\lambda = 3$, $\psi = 0.75$). Dotted lines starting from bottom are t_1, t_2, t_3.

Coalescence behavior at each t_k is determined by the following:

(1) Each ancestral line is part of the big family with probability ψ.
(2) Of the ancestral lines in the big family, each has an equal probability of coalescing into four different "buckets".

This characterization follows from the discrete case. Each child gene is inherited uniformly at random from the available parent genes, so if a child is in a big family, one of their genes is equally likely to be inherited from any of the 2 parents' 4 genes. Each "bucket" corresponds to a gene (see Figure 4).

More rigorously, we can denote the number of lineages entering (not exiting) each t_k as n_k. For example, in Figure 3, $n_1 = 15$. Then coalescence behavior can also be described as such:

(1) The number of lines that are part of the big family at a point t_k is distributed according to $m \sim \text{Binom}(n_k, \psi)$.

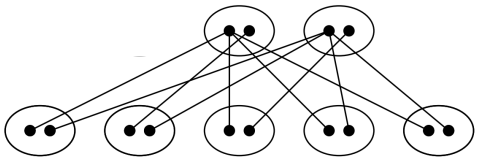


FIGURE 4. Each offspring gene is inherited randomly from 4 parent genes.

(2) Of the m lines, $(m_1, m_2, m_3, m_4) \sim \text{Mn}(m, (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}))$, where Mn denotes the multinomial distribution. Then, m_i lines coalesce together for $i = 1, \cdots, 4$.

Note that under this behavior, some ancestral lines may not be included in the big family, or they might be the only ancestral line coalescing in a gene "bucket". This is more apparent when there are few remaining lines. For example in the second time point t_2 of Figure 3, where 4 lines pass through the HRE without coalescing.

Additionally, note that we can ignore edge cases such as: ancestral lines coming from the same individual cannot coalesce into the same gene; or ancestral lines coalescing into the same parent but different genes will be unable to coalesce in the next generation. These edge cases become negligible as $N \to \infty$. Throughout this report, we will examine the limiting case exclusively.

2.2. **Simulation algorithm.** According to the coalescence behavior we outlined in the previous section, we simulated an HRE tree with n tips (nodes) through an algorithm outlined below:

(1) Initialize $t_{\text{curr}} \leftarrow 0$. Initialize a list of n nodes where $n_{\text{curr}} \leftarrow n$ is the length of this list. Assume n_{curr} updates with the length of the list.
(2) Sample Δt from $\text{Exp}(\lambda)$.
(3) Until the next HRE time point $t_{\text{next}} \leftarrow t_{\text{curr}} + \Delta t$, coalesce the n_{curr} nodes according to the Kingman process (i.e. each pair of nodes coalesces according to an exponential with rate $\frac{1}{2}$).
(4) At t_{next}, sample m from $\text{Binom}(n_{\text{curr}}, \psi)$, where n_{curr} is the number of nodes left.
(5) Sample $(m_1, \cdots, m_4)$ from $\text{Mn}(m, (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}))$.
(6) Permute the existing n_{curr} nodes. Coalesce the first m_1 nodes, then the next m_2 nodes, etc.
(7) Repeat steps 2 through 6 until there is only 1 node left.

Permuting the list of nodes to coalesce randomizes which ancestral lines are chosen. This is assuming *panmixia* (uniform random fertilization).

2.3. **Notation.** It is important to differentiate between the HRE coalescent and the Kingman coalescent. Let H_n be the height of a HRE tree with n tips. Let $H_n^{(\kappa)}$ be the height of a Kingman tree with n tips, where both are random variables. Then $\mathbb{E}[H_n]$ is the expected height averaging over all $\vec{t}$, while $\mathbb{E}^{\vec{t}}[H_n]$ is the expected height conditioned on some fixed $\vec{t}$. Note that $\mathbb{E}[H_n] < \mathbb{E}[H_n^{(\kappa)}]$ and $\mathbb{E}^{\vec{t}}[H_n] < \mathbb{E}[H_n^{(\kappa)}]$ for any $\vec{t}$. HREs only speed up the process of coalescence, never slowing it down.

The Kingman process here is the limiting case of the diploid Wright-Fisher model. As it is the diploid case, each pair of nodes coalesces with rate $\frac{1}{2}$ (as opposed to rate 1 in the haploid model). Thus, time to coalesce from k to $k-1$ nodes is distributed as $T_k \sim \text{Exp}(\frac{1}{2} \cdot \binom{k}{2})$. Then

$$\mathbb{E}[H_n^{(\kappa)}] = \mathbb{E}\left[\sum_{k=2}^{n} T_k\right] = \sum_{k=2}^{n} \mathbb{E}[T_k] = \sum_{k=2}^{n} \frac{4}{k(k-1)} = \boxed{4\left(1 - \frac{1}{n}\right)}.$$

In expectation, the height of any coalescent tree is always less than 4.

In addition, it is necessary to introduce another notion of expectation $\mathbf{E}[\cdot]$, which denotes the average with respect to randomness of the big family times. $\mathbb{E}^{\vec{t}}[\cdot]$ averages with respect to randomness of the coalescent tree *while fixing the big family times*. Finally, $\mathbf{E}[\mathbb{E}^{\vec{t}}[\cdot]]$ averages over both the randomness of the coalescent tree *and* the randomness of the big family times.

3. Results

We wish to study the conditional coalescent under the HRE model. This involves conditioning on the pedigree, i.e. conditioning on the Poisson process. However, conditioning on an infinite-dimensional vector $\vec{t}$ is not well-defined, so we wish to restrict the dimension.

This begs two questions:

(1) How many Poisson points t_i do we need to sufficiently estimate H_n?
(2) Given we only need k Poisson points, can we recover these k points from expected H_n?

With respect to (2), we hope to be able to recover H_n from data, from which we hope to recover the population history **in terms of** $(t_1, \cdots, t_k)$.

3.1. Proving convergence.

Below is a formulation and proof of the convergence of the expected height of the tree conditioned on k time points to the expected height conditioned on all time points.

Lemma 1 (Convergence)**.** *Generate $\vec{t} = (t_1, t_2, \cdots)$ through a Poisson process with rate λ, where $t_1 < t_2 < \cdots$. Fix $\vec{t}$. Then*

$$h_k := \mathbb{E}^{t_1,\cdots,t_k}[H_n]$$

converges as $k \to \infty$. Define its limit to be $h_\infty := \mathbb{E}^{\vec{t}}[H_n]$.

Proof. We assume that $t_1 < t_2 < \cdots$, and that $t_k \xrightarrow{k\to\infty} \infty$.

Recall that $\{h_k\}$ is Cauchy if $\forall \varepsilon > 0, \exists N$ s.t. $\forall j, k \geq N$, $|h_j - h_k| < \epsilon$. WLOG take $j \leq k$. Then we have:

$$h_j = \mathbb{P}^{t_1,\cdots,t_j}(H_n \leq t_j) \cdot \mathbb{E}^{t_1,\cdots,t_j}[H_n | H_n \leq t_j] + \mathbb{P}^{t_1,\cdots,t_j}(H_n > t_j) \cdot \mathbb{E}^{t_1,\cdots,t_j}[H_n | H_n > t_j]$$
$$h_k = \mathbb{P}^{t_1,\cdots,t_j}(H_n \leq t_j) \cdot \mathbb{E}^{t_1,\cdots,t_j}[H_n | H_n \leq t_j] + \mathbb{P}^{t_1,\cdots,t_j}(H_n > t_j) \cdot \mathbb{E}^{t_1,\cdots,t_k}[H_n | H_n > t_j]$$

Note that if we condition on the first k points in the process, if we only wish to know if the process stops at, before, or after t_j, only the first j points will affect the outcome. This can

be seen in blue. Thus, we have

$$|h_j - h_k| = \mathbb{P}^{t_1,\cdots,t_j}(H_n > t_j) \cdot \left| \mathbb{E}^{t_1,\cdots,t_j}[H_n | H_n > t_j] - \mathbb{E}^{t_1,\cdots,t_k}[H_n | H_n > t_j] \right| \tag{1}$$

By Markov's inequality,

$$\mathbb{P}^{t_1,\cdots,t_j}(H_n > t_j) \leq \frac{\mathbb{E}^{t_1,\cdots,t_j}[H_n]}{t_j} < \frac{\mathbb{E}[H_n^{(\kappa)}]}{t_j} = \frac{4(1-\frac{1}{n})}{t_j}$$

Now consider the second factor in (1). Suppose the number of ancestral lines left at time t_j is n_j. Then by the Markov property, the remaining height of the tree is less than the height of a Kingman tree in expectation, *regardless* if we are conditioning on $k-j$ more tips. $\mathbb{E}[H_{n_j}] < \mathbb{E}[H_{n_j}^{(\kappa)}] = 4(1-\frac{1}{n_j}) \leq 4(1-\frac{1}{n})$ for all $n_j \leq n$.

Thus, $\mathbb{E}^{t_1,\cdots,t_j}[H_n | H_n > t_j]$ and $\mathbb{E}^{t_1,\cdots,t_k}[H_n | H_n > t_j]$ must both be contained in the interval $[t_j, t_j + 2(1-\frac{1}{n}))$. Back to (1), we now have

$$|h_j - h_k| < \frac{4(1-\frac{1}{n})}{t_j} \cdot 4\left(1 - \frac{1}{n}\right) = \frac{16(1-\frac{1}{n})^2}{t_j}$$

As t_j goes to infinity, $|h_j - h_k|$ goes to 0. Thus, the sequence is Cauchy and converges. □

Lemma 1 proves convergence in the quenched case. The corollary below proves convergence in the annealed case, i.e. when we do not condition on a specific $\vec{t}$. We will add a superscript to $h_k^{\vec{t}}$ to clarify its dependence on $\vec{t}$. We also use $\mathbb{E}[\cdot]$ to average over all $\vec{t}$.

Corollary 1. $\mathbb{E}[|h_k^{\vec{t}} - h_\infty^{\vec{t}}|]$ *converges to 0 as* $k \to \infty$. *Thus,* $h_k^{\vec{t}}$ *converges in* L^1 *to* $h_\infty^{\vec{t}}$.

Proof. $\mathbb{E}[|h_k^{\vec{t}} - h_\infty^{\vec{t}}|] < \mathbb{E}[\frac{16(1-\frac{1}{n})^2}{t_k}]$ by Lemma 1 and monotonicity of x on the domain $(0, \infty]$. On average, over all vectors $\vec{t}$ where each vector is generated through a Poisson point process with rate λ, t_k is distributed according to $\Gamma(k, \lambda)$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{t_k}\right] &= \int_0^\infty \frac{1}{x} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}\, dx \\
&= \frac{\Gamma(k-1)}{\Gamma(k)} \lambda \int_0^\infty \frac{\lambda^{k-1} x^{k-2} e^{-\lambda x}}{\Gamma(k-1)}\, dx \\
&= \frac{\lambda}{k-1}
\end{aligned}$$

Notice the second integral is the integral of the pdf of a $\Gamma(k-1, \lambda)$ random variable, which evaluates to 1.

Thus, we have that $\mathbb{E}[|h_k^{\vec{t}} - h_\infty^{\vec{t}}|] < \frac{16\lambda(1-\frac{1}{n})^2}{k-1}$. Clearly, this converges to 0 as k goes to infinity. Thus, $h_k^{\vec{t}}$ converges in L^1 to $h_\infty^{\vec{t}}$. □

In summary, we have the bounds:

$$|h_k^{\vec{t}} - h_\infty^{\vec{t}}| < \frac{16(1-\frac{1}{n})^2}{t_j}$$

$$\mathbb{E}[|h_k^{\vec{t}} - h_\infty^{\vec{t}}|] < \frac{16\lambda(1-\frac{1}{n})^2}{k-1}$$

3.2. **Tightening the bound.** Now that we have proved convergence, how many time points is enough to keep h_k and h_∞ sufficiently close? In other words, how large should we take k to bound the error between h_k and h_∞?

Define $\tau_\varepsilon := \inf\{k : \mathbb{E}[|h_k^{\vec{t}} - h_\infty^{\vec{t}}|] < \varepsilon\}$. This value determines the first k for which the error in the annealed case is below ε. If we can bound the τ_ε by some value below, we can answer that question.

Our bounds are quite weak currently. This is primarily due to our use of Markov's inequality in Lemma 1. We present a direct calculation using Tavaré's $g_{nm}(t)$ that tightens the bound in the annealed case.

Lemma 2 (Tavaré bound)**.**

$$\mathbb{E}[|h_k^{\vec{t}} - h_\infty^{\vec{t}}|] < 12\,\frac{n+1}{n}\left(\frac{2\lambda}{2\lambda+1}\right)^j$$

Proof. We specifically wish to tighten the bound on $\mathbb{P}^{t_1,\cdots,t_k}(H_n < t_k)$, which was previously bounded with Markov's inequality.

Fix $t_j \geq 0$. Clearly, $H_n > t_j$ if and only if all n lineages have not coalesced into one lineage by time t_j, so by definition of the functions $g_{nm}(t)$ of (Tav84, Eq. 6.2),

$$\begin{aligned}\mathbb{P}^{t_1,\cdots,t_j}(H_n > t_j) &< 1 - g_{n1}(t_j/2) \\ &= \sum_{k=2}^{n} e^{-k(k-1)t_j/4}\frac{(-1)^k(2k-1)n_{(k)}}{n^{(k)}} \qquad (2)\end{aligned}$$

where $n_{(k)} = n(n-1)\cdots(n-k+1)$, $n^{(k)} = n(n+1)\cdots(n+k-1)$ denote the falling/rising factorials respectively. Here, time is scaled by $1/2$ as we are using the diploid case of Kingman. Now, it can be easily verified that (2) is an alternating sum with terms decreasing in magnitude, so,

$$\left|\sum_{k=2}^{n} e^{-k(k-1)t_j/4}\frac{(-1)^k(2k-1)n_{(k)}}{n^{(k)}}\right| < \left|e^{-t_j/2}\frac{3(-1)^2 n_{(2)}}{n^{(2)}}\right| = 3e^{-t_j/2}\frac{n+1}{n-1}$$

Since under the big family model, $t_j \sim \text{Gamma}(j,\lambda)$, we have

$$\mathbb{E}[\mathbb{P}^{t_1,\cdots,t_j}(H_n > t_j)] < \mathbb{E}\left[3e^{-t_j/2}\frac{n+1}{n-1}\right] = 3\,\frac{n+1}{n-1}\left(\frac{2\lambda}{2\lambda+1}\right)^j. \qquad (3)$$

Now, recall equation (1) of Lemma 1. Taking the expectation over all $\vec{t}$, we have

$$|h_j^{\vec{t}} - h_k^{\vec{t}}| = \mathbb{P}^{t_1,\cdots,t_j}(H_n > t_j)\cdot\left|\mathbb{E}^{t_1,\cdots,t_j}[H_n|H_n > t_j] - \mathbb{E}^{t_1,\cdots,t_k}[H_n|H_n > t_j]\right|$$

$$\begin{aligned}\mathbb{E}[|h_j^{\vec{t}} - h_k^{\vec{t}}|] &< 3\,\frac{n+1}{n-1}\left(\frac{2\lambda}{2\lambda+1}\right)^j \cdot 4(1-\frac{1}{n}) \\ &= 12\,\frac{n+1}{n}\left(\frac{2\lambda}{2\lambda+1}\right)^j.\end{aligned}$$

□

This bound may be poor for small j if λ is large. To get around this, call the rightmost expression of (3) $C_{n,j}$, and look for nonnegative constants $c, \tilde{c}$ with $c + \tilde{c} = 1$ such that $cC_{n,j} + \tilde{c}\tilde{C}_{n,j}$ is minimal, where $\tilde{C}_{n,j}$ is any other upper bound on $\mathbb{E}[\mathbb{P}(H_n > t_j)]$. For instance, $\tilde{C}_{n,j}$ may be taken to be the one derived from applying Markov's inequality to $\mathbb{P}(H_n > t_j)$.

3.3. **Finding an expression for height.** We wish to find an expression for $\mathbb{E}^{t_1,\cdots t_k}[H_n]$. In particular, we will examine the $n = 2$ case.

From (DFBW24, Lemma 1), we have $H_2 \sim \text{Exp}(\frac{1}{2} + \lambda\frac{\psi^2}{4})$ in the annealed case, i.e. not fixing a specific $\vec{t}$.

Lemma 3 (Expectation of H_2, annealed)**.** *Let $0 \leq \psi \leq 1$. Let H_2 be the height of a HRE tree with $n = 2$ tips, where $\vec{t}$ is determined by a Poisson process with rate $\lambda > 0$. Then*

$$\mathbb{E}[\mathbb{E}^{\vec{t}}[H_2]] = \frac{1}{\frac{1}{2} + \lambda\frac{\psi^2}{4}}$$

Proof. Let $\mathbb{P}^{\vec{t}}(H_2 > t)$ be the conditional, limiting, complementary cumulative distribution function as found in (DFBW24, Theorem 1). Note that $Y(t)$ is a step function, representing how many Poisson process points have occurred up to time t. $Y(t)$ is constant on the interval (t_i, t_{i+1}) for all i.

$$\begin{aligned}
\mathbb{E}^{\vec{t}}[H_2] &= \int_0^\infty \mathbb{P}^{\vec{t}}(H_2 > t)dt \\
&= \int_0^\infty e^{-t/2}(1 - \frac{\psi^2}{4})^{Y(t)}dt \\
&= \sum_{i=0}^{\infty} -2e^{-t/2}\left(1 - \frac{\psi^2}{4}\right)^i \Big|_{t_i}^{t_{i+1}} \\
&= \sum_{i=0}^{\infty} -2\left(1 - \frac{\psi^2}{4}\right)^i \left(e^{-t_{i+1}/2} - e^{-t_i/2}\right) \qquad (4)
\end{aligned}$$

As $\vec{t}$ is generated by a Poisson process with rate λ, $t_i \sim \Gamma(i, \lambda)$. Let M_X be the moment generating function for a random variable distributed according to X.

Averaging over all $\vec{t}$ in (4) gives

$$\begin{aligned}
\mathbb{E}[\mathbb{E}^{\vec{t}}[H_2]] &= \sum_{i=0}^{\infty} -2\left(1 - \frac{\psi^2}{4}\right)^i \left(\mathbb{E}[e^{-t_{i+1}/2}] - \mathbb{E}[e^{-t_i/2}]\right) \\
&= \sum_{i=0}^{\infty} -2\left(1 - \frac{\psi^2}{4}\right)^i \left(M_{\Gamma(i+1,\lambda)}(-1/2) - M_{\Gamma(i,\lambda)}(-1/2)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=0}^{\infty} -2\left(1-\frac{\psi^2}{4}\right)^i \left(\left(1+\frac{1}{2\lambda}\right)^{-(i+1)} - \left(1+\frac{1}{2\lambda}\right)^{-i}\right) \\
&= \sum_{i=0}^{\infty} -2\left(1-\frac{\psi^2}{4}\right)^i \left(-\frac{1}{2\lambda}\right)\left(1+\frac{1}{2\lambda}\right)^{-(i+1)} \\
&= \frac{1}{\lambda(1+\frac{1}{2\lambda})} \sum_{i=0}^{\infty} \left(\frac{1-\frac{\psi^2}{4}}{1+\frac{1}{2\lambda}}\right)^i \\
&= \frac{1}{\lambda+1/2} \cdot \frac{1}{\frac{(1+\frac{1}{2\lambda})-(1-\frac{\psi^2}{4})}{1+\frac{1}{2\lambda}}} \\
&= \boxed{\frac{1}{\frac{1}{2}+\lambda\frac{\psi^2}{4}}}.
\end{aligned}
$$

This completes the proof. □

Lemma 4 (Expectation of H_2, conditioned on the first k Poisson points)**.** *Let $0 \le \psi \le 1$. Let H_2 be the height of a HRE tree with $n = 2$ tips, where $\vec{t}$ is determined by a Poisson process with rate $\lambda > 0$. Then $\mathbb{E}^{t_1,\cdots,t_k}[H_2]$ is a linear combination of $\{e^{-t_i/2}\}_{i=0,\cdots,k}$.*

Proof. Define $\vec{t'} = (t_{k+1}-t_k, t_{k+2}-t_k, \cdots)$. Note $\vec{t'}$ is distributed identically to $\vec{t}$, i.e. according to a Poisson process with rate λ. We rearrange the first sum in the calculation below.

$$
\begin{aligned}
E^{t_1,\cdots,t_k}[H_2] &= \sum_{i=0}^{k-1} -2\left(1-\frac{\psi^2}{4}\right)^i \left(e^{-t_{i+1}/2} - e^{-t_i/2}\right) \\
&\quad + E\left[\sum_{i=k}^{\infty} -2\left(1-\frac{\psi^2}{4}\right)^i \left(e^{-t_{i+1}/2} - e^{-t_i/2}\right)\right] \\
&= -2\sum_{i=0}^{k-1}\left(1-\frac{\psi^2}{4}\right)^i \left(e^{-t_{i+1}/2} - e^{-t_i/2}\right) + \left(1-\frac{\psi^2}{4}\right)^k e^{-t_k/2} \cdot E[E^{\vec{t'}}[H_2]] \\
&= 2 - \frac{\psi^2}{2}\sum_{i=1}^{k-1}\left(1-\frac{\psi^2}{4}\right)^{i-1} e^{-t_i/2} - 2\left(1-\frac{\psi^2}{4}\right)^{k-1} e^{-t_k/2} + \frac{(1-\frac{\psi^2}{4})^k}{\frac{1}{2}+\lambda\frac{\psi^2}{4}} e^{-t_k/2} \\
&= \boxed{2 - \frac{\psi^2}{2}\left[\sum_{i=1}^{k-1}\left(1-\frac{\psi^2}{4}\right)^{i-1} e^{-t_i/2}\right] - \frac{(2\lambda+1)\frac{\psi^2}{4}}{\frac{1}{2}+\lambda\frac{\psi^2}{4}}\left(1-\frac{\psi^2}{4}\right)^{k-1} e^{-t_k/2}}
\end{aligned}
$$

□

The calculation methods in Lemma 3, Lemma 4 could be used to find expectations for $\phi(H_2)$ for any ϕ continuous, in particular $\phi(t) = e^{-ut}$.

Lemma 5 (Expectation of e^{-uH_2}, annealed)**.** *For any $u \in (0,\infty)$, $0 \le \psi \le 1$, H_2 and $\vec{t}$ defined as before, then*

$$
\mathbb{E}[\mathbb{E}^{\vec{t}}[e^{-uH_2}]] = \frac{\frac{1}{2}+\lambda\frac{\psi^2}{4}}{\frac{1}{2}+\lambda\frac{\psi^2}{4}+u}
$$

Proof. Proof similar to Lemma 3; see Appendix. □

Lemma 6 (Expectation of e^{-uH_2}, conditioned on first k Poisson points)**.** *For any $u \in (0, \infty)$, $0 \leq \psi \leq 1$, H_2 and $\vec{t}$ defined as before, then $\mathbb{E}^{t_1,\cdots,t_k}[e^{-uH_2}]$ is a linear combination of $\{e^{-t_i(1/2-u)}\}_{i=0,\cdots,k}$.*

Proof. Proof similar to Lemma 4; see Appendix. The expression is

$$\mathbb{E}^{t_1,\cdots,t_k}[e^{-uH_2}] = 1 - \frac{u}{u+1/2} + \frac{u}{u+1/2} \cdot \frac{\psi^2}{4} \sum_{i=1}^{k-1} \left(1 - \frac{\psi^2}{4}\right)^{i-1} e^{t_i(-1/2-u)}$$
$$+ \frac{u}{u+1/2} \cdot \frac{1/2 + (\lambda+u)\frac{\psi^2}{4}}{1/2 + \lambda\frac{\psi^2}{4} + u} \left(1 - \frac{\psi^2}{4}\right)^{k-1} e^{t_k(-1/2-u)}$$

□

3.4. **Proving statistical identifiability of an estimator.** In future work, we hope to use Lemmas 3-6 to estimate $t_1, \cdots, t_k$ for some k.

4. Discussion

In this report, the main object of interest was the height of the coalescent tree. However, there are other statistics from which we can extract additional information. One of these statistics is the site frequency spectrum.

4.1. **Site frequency spectrum.** The site frequency spectrum (SFS) is a summary statistic that aligns with the infinite sites model of mutation.

The **infinite sites model** is a mathematical model of mutation such that a mutation happens at each site in a DNA sequence at most once (i.e. there is no recombination) and each site has at most two different nucleotides. These assumptions can be interpreted as the evolution of long DNA sequences with very low mutation rate, such that each nucleotide can mutate at most once.

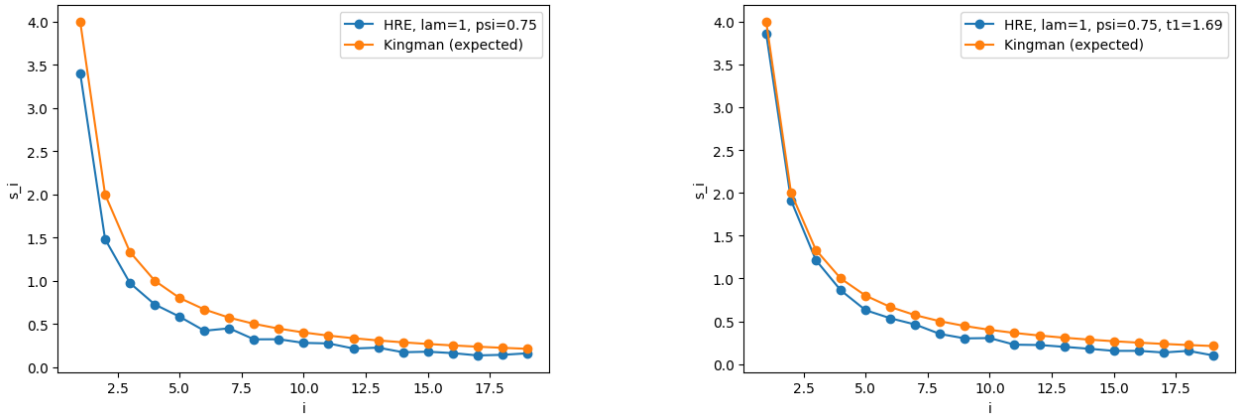
Assuming we know the ancestral allele at each polymorphic site (a site where there is more than one nucleotide, also known as a **segregating site**), we can count the number of sites ξ_i where there are i mutant alleles and $n - i$ ancestral alleles. These ξ_i can then be plotted in the site frequency spectrum. This case, where we know the ancestral allele is called the **unfolded** SFS, while the opposite case is called the **folded** SFS. If we do not know the ancestral allele, we simply count the number of minor alleles at each polymorphic site. Then we have $\eta_i = \frac{\xi_i + \xi_{n-i}}{1+\delta_{i,n-i}}$ for $1 \leq i \leq [n/2]$.

Two other statistics of note are S, the number of segregating sites in a sample, and π, the average difference between pairs of sequences in a sample. Note that $S = \sum_{i=1}^{n-1} \xi_i$, the sum of the values in an SFS. $\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{ij}$, where k_{ij} is the number of differences between sequence i and j. The SFS can be thought of as an intermediate between S and π.

4.2. **Branch frequency spectrum.** For clarity, I will refer to the spectrum obtained from a coalescent tree as the branch frequency spectrum, although it is often referred to as an SFS as well.

The branch frequency spectrum is calculated as such: each s_i is the sum of the lengths of each branch that subtends i tips. Under the infinite sites model, mutations occurs on a branch at a rate of $\theta/2$ per unit length. Any mutation that occurs on a branch that subtends i tips will manifest itself in exactly i sequences. Thus, the number of sites which have i mutant alleles will be proportional to the sum of the lengths of the branches that subtend i tips. In addition, the total length of the tree is proportional to the number of total mutations, which is equal to the number of segregating sites S under the infinite sites model.

In this way, the branch frequency spectrum is related to the site frequency spectrum, and both statistics summarize the same information but from different starting points.



(A) Unconditional BFS

(B) Conditional BFS

FIGURE 5. Branch frequency spectra for HRE vs. Kingman coalescents, $n = 20$ (each plot shows s_i vs. i)

Figure 5a shows in blue the average of branch frequency spectra over all Poisson processes $\vec{t}$ with $\lambda = 1, \psi = 0.75$. Figure 5b generates and fixes a specific Poisson process with the same parameters, specifically $\vec{t} = [0.28, 0.54, 0.95, 1.97, \cdots]$. One can see that the values in 5b are closer to the expected Kingman spectrum values. In both spectra, the HRE spectrum is lower in value at all points compared to the expected Kingman spectrum. This is reasonable, as an HRE tree will always coalesce more quickly than a Kingman tree.

In future work, these s_i can be examined as opposed to H_n.

5. Conclusion

In this report, we explored the multiple merger coalescent model, focusing on "big family" events and their impact on the coalescent tree structure. Specifically, we aimed to understand how much population history is necessary to accurately approximate the time to the most recent common ancestor (MRCA), as well as investigate the possibility of recovering population histories from genomic data.

We addressed two core questions:

(1) **How much of the population history do we need to accurately examine time to MRCA?** Specifically, how many Poisson points of $\vec{t}$ do we need to approximate H_n up to some error?
(2) **Can we recover population histories from given data?** Specifically, if we only need k Poisson points, can we recover these k points from expected H_n?

We made significant progress on question (1) and the beginnings of question (2). Future work includes:

- Tightening bounds to reduce the number of time points required.
- Proving statistical identifiability of an estimator for time points t_i.
- Applying this estimator to real-world data.
- Examining other statistics, such as the branch frequency spectrum.

In conclusion, we have made promising progress toward understanding the genealogical structure of populations subject to "big family" events. The insights gained from this work pave the way for more accurate reconstructions of population history and the development of tools to detect and analyze significant demographic events. While there is still much to explore, particularly in the realm of statistical identifiability and real-world application, the framework we have established provides a solid foundation for future research in this area.

Appendix

Proof of Lemma 5.

Proof.

$$\begin{aligned}
\mathbb{E}^{\vec{t}}[e^{-uH_2}] &= \int_0^\infty P^{\vec{t}}(e^{-uH_2} > t)dt \\
&= \int_0^1 P^{\vec{t}}(H_2 < \frac{-\ln t}{u})dt \\
&= \int_0^1 1 - \left(e^{\frac{\ln t}{2u}}\left(1-\frac{\psi^2}{4}\right)^{Y(\frac{-\ln t}{u})}\right)dt \\
&= 1 - \int_0^1 t^{\frac{1}{2u}}\left(1-\frac{\psi^2}{4}\right)^{Y(\frac{-\ln t}{u})}dt \\
&= 1 - \sum_{k=0}^\infty \frac{u}{u+1/2} t^{\frac{1}{2u}+1}\left(1-\frac{\psi^2}{4}\right)^k \Big|_{t=e^{-ut_{k+1}}}^{t=e^{-ut_k}} \\
&= 1 - \frac{u}{u+1/2}\sum_{k=0}^\infty \left(1-\frac{\psi^2}{4}\right)^k \left(e^{t_k(-1/2-u)} - e^{t_{k+1}(-1/2-u)}\right)
\end{aligned}$$

Taking the expectation of this, we can plug in $-1/2-u$ into the gamma moment generating functions. We have

$$\begin{aligned}
\mathbb{E}[\mathbb{E}^{\vec{t}}[e^{-uH_2}]] &= 1 - \frac{u}{u+1/2}\sum_{k=0}^\infty \left(1-\frac{\psi^2}{4}\right)^k \left(\left(1+\frac{1/2+u}{\lambda}\right)^{-k} - \left(1+\frac{1/2+u}{\lambda}\right)^{-(k+1)}\right) \\
&= \boxed{\frac{\frac{1}{2}+\lambda\frac{\psi^2}{4}}{\frac{1}{2}+\lambda\frac{\psi^2}{4}+u}}
\end{aligned}$$

Summing over the resulting geometric series, we have the moment generating function of an exponential random variable with rate $\frac{1}{2}+\lambda\frac{\psi^2}{4}$, consistent with (DFBW24, Lemma 1).

□

Proof of Lemma 6.

Proof.

$$\mathbb{E}^{t_1,\cdots,t_k}[e^{-uH_2}] = 1 - \frac{u}{u+1/2}\sum_{i=0}^{k-1}\left(1-\frac{\psi^2}{4}\right)^i \left(e^{t_i(-1/2-u)} - e^{t_{i+1}(-1/2-u)}\right) \tag{5}$$

$$- \mathbb{E}\left[\frac{u}{u+1/2}\sum_{i=k}^{\infty}\left(1-\frac{\psi^2}{4}\right)^i \left(e^{t_i(-1/2-u)} - e^{t_{i+1}(-1/2-u)}\right)\right] \tag{6}$$

Rearranging the top expression, we have

$$(5) = 1 - \frac{u}{u+1/2} + \frac{u}{u+1/2} \cdot \frac{\psi^2}{4} \sum_{i=1}^{k-1} \left(1 - \frac{\psi^2}{4}\right)^{i-1} e^{t_i(-1/2-u)} + \frac{u}{u+1/2} \left(1 - \frac{\psi^2}{4}\right)^{k-1} e^{t_k(-1/2-u)}$$

Evaluating the bottom expression, we have

$$\begin{aligned}(6) &= -\frac{u}{u+1/2} \left(1 - \frac{\psi^2}{4}\right)^{k} e^{t_k(-1/2-u)} \mathbf{E}[1 - \mathbb{E}^{\vec{t}}[e^{-uH_2}]] \\ &= -\frac{u}{u+1/2} \left(1 - \frac{\psi^2}{4}\right)^{k} e^{t_k(-1/2-u)} \frac{u}{\frac{1}{2} + \lambda\frac{\psi^2}{4} + u}\end{aligned}$$

Summing the two expressions, we have our final expression

$$(5) + (6) = 1 - \frac{u}{u+1/2} + \frac{u}{u+1/2} \cdot \frac{\psi^2}{4} \sum_{i=1}^{k-1} \left(1 - \frac{\psi^2}{4}\right)^{i-1} e^{t_i(-1/2-u)} + \frac{u}{u+1/2} \cdot \frac{1/2 + (\lambda+u)\frac{\psi^2}{4}}{1/2 + \lambda\frac{\psi^2}{4} + u} \left(1 - \frac{\psi^2}{4}\right)^{k-1} e^{t_k(-1/2-u)}$$

□

Acknowledgements

This report is based on work supported by NSF grant DMS-2051032, which I gratefully acknowledge. I would also like to express my thanks to the Mathematics department of Indiana University for hosting the program, my mentor Wai-Tong (Louis) Fan, as well as Dimitrios Diamantidis and Daniel Rickert, who all aided me significantly in my research.

References

[DFBW24] D. Diamantidis, W.L. Fan, M. Birkner, and J. Wakeley, *Bursts of coalescence within population pedigrees whenever big families occur*, Genetics **227** (2024), no. 1, iyae030.

[Tav84] S. Tavaré, *Line-of-descent and genealogical processes, and their applications in population genetics models*, Theoretical Population Biology **26** (1984), 119–164.

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA

Email address: aliciaz@mit.edu