**Approximate sampling formulas under the coalescent with finite-alleles models of mutation**

by

Anand Bhaskar

A thesis submitted in partial satisfaction of the
requirements for the degree of
Master of Arts

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair
Professor Steven Evans
Professor Montgomery Slatkin

Fall 2012

**Approximate sampling formulas under the coalescent with finite-alleles models of mutation**

## Abstract

Approximate sampling formulas under the coalescent with finite-alleles models of mutation

by

Anand Bhaskar

Master of Arts in Statistics

University of California, Berkeley

Professor Yun S. Song, Chair

In population genetics, the coalescent is a family of models that is widely used to describe the ancestry of a sample of chromosomes taken from a large randomly mating population. Many applications in population genetics utilize sampling distributions, which describe the probability of generating a given sample under the coalescent. For a single locus with special models of mutation such as the infinite-alleles or the finite-alleles parent-independent mutation model, closed-form expressions for the sampling probability have been known for many decades. However, no exact formula is currently known for more general models of mutation that are of biological interest.

In this thesis, we review known closed-form results about the one-locus sampling distribution for the infinite-alleles and finite-alleles parent-independent mutation models, and proceed to derive novel first-order approximate closed-form sampling formulas for general finite-alleles mutation models. More precisely, we give efficiently evaluable approximate sampling formulas for an arbitrary irreducible recurrent mutation model or for a reversible recurrent mutation model, depending on whether the number of distinct observed allele types in the sample is at most three or four, respectively. These results are derived by applying combinatorial identities to an urn construction related to the coalescent. We also give alternate proofs for several of our results using martingale and coupling arguments. We demonstrate using numerical computations that our first-order approximate formulas are highly accurate compared to the exact sampling probabilities when the per-base mutation rate is low, which holds in many settings of biological interest.

*To the memory of my grandmother*

# Contents

## Acknowledgments

I am very thankful to my advisor Yun Song for being the most supportive and encouraging mentor that a graduate student could expect. Yun has always given me free rein to work on the problems that I find interesting, and collaborating with him has been one of the biggest joys of my graduate school experience. I am also very grateful to my good friend and collaborator, Jack Kamm, with whom I spent a most enjoyable summer working on the research described in this thesis. Yun's research group has also been a very friendly and welcoming environment for exchanging research ideas, and I gratefully acknowledge all my fellow students and postdocs. I would particularly like to thank Andrew Chan, Paul Jenkins, Joshua Paul and Matthias Steinrücken for their collegiality and discussions from which I have greatly benefitted. I also wish to thank Monty Slatkin and Steve Evans for many interesting discussions, suggestions and sage advice over the years, and also for serving on my thesis committee.

My time in graduate school would not have been nearly as enjoyable were it not for the camaraderie of my most wonderful friends and office mates of many years, Siu Man Chan, Siu On Chan and Thomas Watson, and a group of close friends in the computer science theory group, especially James Cook, Anindya De, Kevin Dick and Piyush Srivastava. I would be remiss if I did not acknowledge La Shana Porlaris, who has always cheerfully helped me with administrativia, both in the EECS and in the Statistics departments. I am also grateful to the generous support provided by the Berkeley Fellowship during my initial years of graduate school. Finally and most importantly, words cannot do justice to the debt of gratitude I owe my parents, who have always supported and believed in me.

# Chapter 1

# Introduction

Population genetics is a quantitative subfield of evolutionary biology that is concerned with the evolution of genetic variation in a population over time. The locations of the genome which exhibit genetic variation are called *loci* (sing. *locus*), and the variations at these loci are called *alleles*. For example, a locus could be a single nucleotide of DNA, with the set of alleles being the four possible DNA nucleotides. The distribution of the alleles in a population fluctuates over time due to the biological processes of mating, mutation, recombination and natural selection, among other things. The most basic model of population evolution, the *Wright-Fisher* model [7, 46], describes the alleles at a single locus of a population of size $N$ as evolving in discrete generations, where the alleles in each generation are generated by randomly sampling parent alleles from the previous generation while giving them an opportunity to mutate. The Wright-Fisher model thus describes the forward-time dynamics of the entire population. Under a suitable rescaling of parameters and time, one can create a continuous-time Markov process called the *coalescent* [25, 26, 24] that can approximate the dynamics of the discrete-time Wright-Fisher process, while allowing one to more easily restrict attention to the genealogical history *backwards in time* of a *finite sample* of individuals drawn from the population. Since its introduction, the coalescent has become a mainstay of modern population genetics, and has been extended to incorporate time-varying population sizes [13], multiple loci with recombination [9], natural selection [27], and population subdivision [16]. The importance of the coalescent stems from the fact that it is the continuous-time scaling limit of a large class of discrete-time models of random mating [26, 32, 33], and hence robust to the minutiae of these random mating schemes.

## 1.1 Sampling distribution

While the coalescent is a useful mathematical framework for performing model-based full-likelihood analyses, many quantities of interest are intractable to compute under the coalescent for large sample sizes. One such fundamental problem is that of determining the probability of sampling a particular set of individuals who are specified by their sequence of

alleles at one or more loci (called *haplotypes*). This likelihood computation is at the heart of algorithms that infer biological parameters of the population, such as the mutation rate at a locus [42, 4] or the crossover recombination rate between loci [18, 31, 34]. This likelihood is also needed for many important applications like inferring the genealogical ancestry of a sample of individuals [12], phasing genotype data into pairs of haplotype sequences [41], estimating the time a mutation arose in the population [15], imputing missing data at a locus [43, 41], etc.

Even when considering only one locus, obtaining a closed-form formula for the sampling probability has remained elusive for general mutation models. A well-known exception to this complication is the celebrated Ewens' sampling formula (ESF) [6], which describes the probability distribution of a sample configuration under the one-locus infinite-alleles model of mutation. In the case of finitely-many alleles, a closed-form sampling formula is known [47] only for the parent-independent mutation (PIM) model, in which the probability of mutating from allele $j$ to allele $i$ depends only on the child allele $i$. For a general non-PIM mutation model, finding an exact, closed-form sampling formula has remained a challenging open problem.

In this thesis, we describe the one-locus coalescent model for the infinite-alleles and the finite-alleles models of mutation and review known formulas for the sampling probability. We then make progress on the problem of computing one-locus sampling probabilities under general finite-alleles mutation models by deriving approximate, closed-form sampling formulas that asymptotically match the exact sampling probability in the regime of low mutation rates. More precisely, given a sample configuration $\boldsymbol{n}$ and the model parameters (mutation rate $\theta$ and transition matrix $\boldsymbol{P}$), we consider the Taylor expansion of the sampling probability $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$ about $\theta = 0$. As discussed later, if $\boldsymbol{P}$ is irreducible when restricted to the observed alleles in the sample, then the leading order term in the expansion is proportional to $\theta^{|\mathcal{O}_{\boldsymbol{n}}|-1}$, where $|\mathcal{O}_{\boldsymbol{n}}|$ is the number of distinct observed alleles in the sample configuration $\boldsymbol{n}$. Hence,

$$q(\boldsymbol{n} \mid \theta, \boldsymbol{P}) = \theta^{|\mathcal{O}_{\boldsymbol{n}}|-1} Q(\boldsymbol{n} \mid \boldsymbol{P}) + O(\theta^{|\mathcal{O}_{\boldsymbol{n}}|}), \tag{1.1}$$

where $Q(\boldsymbol{n} \mid \boldsymbol{P})$ is the leading order coefficient that depends on the mutation transition matrix $\boldsymbol{P}$ but not on the mutation rate $\theta$. In this thesis, we consider the problem of obtaining exact closed-form formulas for $Q(\boldsymbol{n} \mid \boldsymbol{P})$, which can be used to give asymptotic approximations for $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$. As many organisms typically have small per-base mutation rates, our results are of biological interest.

The rest of this thesis is organized as follows. In Chapter 2, we describe the coalescent model and survey known formulas for the one-locus sampling distribution. We state our main closed-form approximate sampling formulas in Chapter 3, and prove them in Chapter 4. We prove our results by constructing an urn process related to the coalescent, and use this urn process to develop a recursion for the approximate sampling probability. Using combinatorial identities, closed-form solutions to this recursion can be obtained. We also give alternate proofs for some of our formulas using martingale and coupling arguments. Many of the results and proofs in this thesis can also be found in our paper [2]. In Chapter 5, we demonstrate

the applicability of our approximate sampling formulas using biologically realistic mutation matrices. We conclude with a discussion of problems for future research in Chapter 6.

# Chapter 2

# Preliminaries

## 2.1 Notation

We first introduce some notation which will be used throughout this thesis.

**Definition 1** ($\boldsymbol{n}$, sample configuration)**.** *A sample of individuals is denoted by* $\boldsymbol{n} = (n_i)_{i\in[K]}$*, where* $n_i \in \mathbb{Z}_{\geq 0}$ *denotes the number of individuals in the sample with allele* $i$*, and* $1 \leq i \leq K$*. The size* $|\boldsymbol{n}|$ *of the sample* $\boldsymbol{n}$ *is denoted by the same letter in non-bold-face,* $n$*. For notational convenience, we use* $\boldsymbol{e}_i$ *to denote the sample configuration with a single individual of type* $i$ *and write* $\boldsymbol{n} = n_1\boldsymbol{e}_1 + \cdots + n_K\boldsymbol{e}_K$*. For a subset* $S \subseteq [K]$*, we define* $\boldsymbol{n}_S = \sum_{i\in S} n_i\boldsymbol{e}_i$ *and* $n_S = |\boldsymbol{n}_S|$*.*

For a sample configuration $\boldsymbol{n}$, we define the combinatorial quantity $\Lambda(\boldsymbol{n})$ as

$$\Lambda(\boldsymbol{n}) = \frac{\prod_{i:n_i>0}(n_i-1)!}{(n-1)!}. \tag{2.1}$$

For $k \in \mathbb{Z}_{\geq 0}$ and $x \in \mathbb{R}$, the $k$th *falling* factorial of $x$ (denoted $(x)_{k\downarrow}$) and the $k$th *rising* factorial of $x$ (denoted $(x)_{k\uparrow}$) are defined as

$$\begin{aligned}(x)_{k\downarrow} &= x(x-1)\cdots(x-k+1),\\ (x)_{k\uparrow} &= x(x+1)\cdots(x+k-1),\end{aligned}$$

with $(x)_{0\downarrow} = (x)_{0\uparrow} = 1$. The $k$th harmonic number $H_k$ is defined as

$$H_k = 1 + \frac{1}{2} + \cdots + \frac{1}{k},$$

with $H_0 = 0$.

## 2.2 Generative model

In this section, we briefly review the one-locus coalescent model which generates the data in our samples. For a more comprehensive introduction, consult the text by Wakeley [45]. One of the most commonly used models of random mating in population genetics, the Wright-Fisher model [7, 46], considers a population of size $N$ evolving in discrete generations. In each generation, a population of $N$ offsprings is generated, where an offspring chooses its parent uniformly at random from the previous generation. By rescaling the unit of time and taking the population size $N$ to infinity, one can derive a continuous-time Markov process that can be used to trace the genealogical ancestry backwards-in-time of a sample taken from the population at present. Given a sample of size $n$, we can define a continuous-time Markov process, Kingman's $n$-coalescent [25, 26, 24], with state space $\mathcal{P}_{[n]}$, the set of partitions of $\{1, \ldots, n\}$. The integers in $[n]$ index the individuals in our sample, and the occurrence of $i$ and $j$ in the same block of the partition at time $t$ denotes that the sample individuals $i$ and $j$ had the same ancestor at time $t$ in the past. Letting $\{\Pi_n(t), t \geq 0\}$ denote this Markov process, the time evolution of the state $\Pi_n(t)$ is determined by the following parameters:

- Initial condition: $\Pi_n(0) = \{\{1\}, \{2\}, \ldots, \{n\}\}$. This encodes the fact that at the present time ($t = 0$), none of the individuals in the sample share any common ancestors.
- Transitions: Given states $\alpha, \beta \in \mathcal{P}_{[n]}$, the transition rate $q_{\alpha\beta}$ from state $\alpha$ to state $\beta$ is given by,

$$q_{\alpha\beta} = \begin{cases} 1 & \text{if } \alpha \lessdot \beta \\ 0 & \text{otherwise.} \end{cases}$$

  Here, the notation $\alpha \lessdot \beta$ denotes that the partition $\beta$ can be obtained by merging exactly two blocks of the partition $\alpha$. The reason for these rates is as follows: if there are $|\alpha|$ ancestral lineages of the sample individuals at time $t$, then each pair of these ancestral lineages will find a common ancestor (i.e. *coalesce*) after an exponentially distributed amount of time with mean parameter 1. Since each ancestral lineage is represented by a block of $\alpha$, the resulting state after such a coalescence will be a partition $\beta$ having $|\beta| = |\alpha| - 1$ ancestral lineages and where some two blocks of $\alpha$ have merged into a single block in $\beta$.

The above described partition-valued process models the genealogical tree relating a randomly drawn sample of individuals from the present. After drawing a random tree over the $n$ sample individuals according to the $n$-coalescent, mutations are dropped on the tree according to a Poisson point process with rate $\theta/2$. Each mutation changes the allelic type of the lineage on which it falls. By mutating the allelic type of lineages according to the specified mutation model and propagating this allelic type information forwards-in-time starting from the most recent common ancestor of the sample, the allele configuration of the sample at the present time can be generated. We now describe two models of mutation, the infinite-alleles

and the finite-alleles models, which differ in how the mutations change the allelic type of the lineages on which they occur.

### Infinite-alleles model of mutation

In this mutation model, each mutation that occurs on the genealogical tree relating the $n$ sample individuals creates a new allelic type that has never been seen before. The configuration of a sample of size $n$ at the present time is described by $\boldsymbol{n} = (n_1, \ldots, n_K)$, where $K$, $1 \leq K \leq n$ is the number of distinct alleles observed in the sample, and $n_i$ is the number of alleles of type $i$. In this model, $K$ is a random variable that depends on the observed sample configuration $\boldsymbol{n}$.

### Finite-alleles model of mutation

In the finite-alleles mutation model, the allele observed at a locus can be one of $K$ types, where $K$ is a fixed constant. For example, $K = 4$ for a SNP locus, and $K = 20$ for an amino acid locus. The effect of mutations on the allelic types of lineages is specified by an irreducible stochastic matrix $\boldsymbol{P}$. When a mutation occurs on a lineage carrying an allele of type $j$, the mutation changes the allelic type of the lineage to $i$ with probability $P_{ji}$. The allelic type of the most recent common ancestor of the individuals in the sample is randomly drawn from the stationary distribution of this mutation matrix $\boldsymbol{P}$.

## 2.3 Known closed-form results

For the infinite-alleles mutation model and a specific family of finite-alleles mutation models, closed-form formulas are known for the sampling distribution. We will use $q(\boldsymbol{n} \mid \theta)$ and $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$ to denote the sampling probability of any *ordered* sample having configuration $\boldsymbol{n}$ under the infinite-alleles and the finite-alleles mutation models, respectively. By exchangeability of the individuals in the sample, the probability of any ordered sample with configuration $\boldsymbol{n}$ is invariant under all permutations of the sampling order.

### Infinite-alleles model

By considering the above described generative model and conditioning on the most recent genealogical event back in time, one can derive a linear system of equations relating the ordered sampling probability for different sample configurations. Furthermore, the variables in this system of equations have an elegant closed-form formula. This is made precise in the following theorem.

**Theorem 1.** *[6, 5] The sampling probability of an ordered sample configuration $\boldsymbol{n}$ in the infinite-alleles model, $q(\boldsymbol{n} \mid \theta)$, is the unique solution to the following system of equations,*

$$q(\boldsymbol{n} \mid \theta) = \frac{n-1}{n-1+\theta} \sum_{i=1}^{K} \frac{n_i(n_i-1)}{n(n-1)} q(\boldsymbol{n} - \boldsymbol{e}_i \mid \theta) + \frac{\theta}{n-1+\theta} \sum_{i=1}^{K} \frac{\delta_{n_i,1}}{n} q(\boldsymbol{n} - \boldsymbol{e}_i \mid \theta), \quad (2.2)$$

*with boundary conditions $q(\boldsymbol{e}_i \mid \theta) = 1$ for $i = 1, \ldots, K$, and where $K$ is the number of observed alleles in the sample configuration $\boldsymbol{n}$.*

*The solution $q(\boldsymbol{n} \mid \theta)$ to this system of equations and boundary conditions is given by*

$$q(\boldsymbol{n} \mid \theta) = \left( \prod_{i=1}^{K} (n_i - 1)! \right) \frac{\theta^K}{(\theta)_{n\uparrow}}. \quad (2.3)$$

The formula in (2.3) is known as Ewens' sampling formula (ESF) [6]. A Pólya-like urn model interpretation [17] of formula (2.3) has been known for some time, and recently a new combinatorial proof of the ESF has been given [11]. Furthermore, the ESF also arises in several interesting contexts outside biology, including random partition structures; the ESF is a special case of the two-parameter sampling formula [37, 38] for exchangeable random partitions. See [1] for examples of other interesting combinatorial connections.

## Finite-alleles model

Similar to the infinite-alleles model, by conditioning on the most recent genealogical event back in time in the coalescent process, one can derive a system of linear equations between the sampling probabilities $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$ for different sample configurations $\boldsymbol{n}$. We have the following theorem.

**Theorem 2.** *[47, 29, 14] The sampling probability of an ordered sample configuration $\boldsymbol{n}$ in a $K$-alleles mutation model, $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$, is the unique solution to the following system of equations,*

$$\begin{aligned} q(\boldsymbol{n} \mid \theta, \boldsymbol{P}) = {} & \frac{n-1}{n-1+\theta} \sum_{i=1}^{K} \frac{n_i(n_i-1)}{n(n-1)} q(\boldsymbol{n} - \boldsymbol{e}_i \mid \theta, \boldsymbol{P}) \\ & + \frac{\theta}{n-1+\theta} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{n_i}{n} P_{ji} q(\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j \mid \theta, \boldsymbol{P}), \end{aligned} \quad (2.4)$$

*with boundary conditions $q(\boldsymbol{e}_i \mid \theta, \boldsymbol{P}) = \pi_i$ for $i = 1, \ldots, K$, where $\boldsymbol{\pi}$ is the stationary distribution of $\boldsymbol{P}$.*

*Furthermore, if $\boldsymbol{P}$ satisfies $P_{ji} = \pi_i$ for all $1 \leq i, j \leq K$, then the solution to this system of equations is given by*

$$q(\boldsymbol{n} \mid \theta, \boldsymbol{P}) = \frac{\prod_{i=1}^{K} (\theta \pi_i)_{n_i\uparrow}}{(\theta)_{n\uparrow}}. \quad (2.5)$$

The mutation matrices satisfying the conditions $P_{ji} = \pi_i$ for all $1 \leq i, j \leq K$ are called parent-independent mutation (PIM) matrices, and the sampling probability given in (2.5) is also known as Wright's sampling formula (WSF) [47]. Under such mutation models, the probability of mutating from allele $j$ to allele $i$ depends only on the child allele $i$. No closed-form formula for $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$ is known for more general $K$-allele models of mutation.

For a given sample configuration $\boldsymbol{n}$ and fixed parameters $\theta$ and $\boldsymbol{P}$, the system of equations (2.4) has $\binom{n+K}{K} - 1 = O(n^K)$ variables. Even for moderate sample sizes of a few hundred individuals and $K = 4$ as in the case of SNP loci, solving such a system of equations imposes a prohibitively high computational cost, especially since this probability computation would typically be repeated in each iteration of a Markov chain Monte-Carlo (MCMC) algorithm. One of the main motivations for our closed-form approximate sampling distributions is to reduce this high computational complexity. Using some precomputation, our formulas can be evaluated in constant time and space.

## 2.4 Closed-form approximations

In the rest of this thesis, we will work with Kingman's coalescent with a $K$-allelic recurrent mutation model specified by the population-scaled mutation rate $\theta/2$ and ergodic mutation transition matrix $\boldsymbol{P}$, where $P_{ji}$ denotes the probability of allele $j$ mutating to allele $i$ forward in time given that a mutation occurs. The stationary distribution of $\boldsymbol{P}$ is denoted by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$.

The following definitions will be used throughout the rest of this thesis:

**Definition 2** ($\mathcal{O}_{\boldsymbol{n}}$, observed allele types)**.** *Given a sample $\boldsymbol{n}$, let $\mathcal{O}_{\boldsymbol{n}} \subseteq [K]$ denote the set of observed allele types; i.e., $\mathcal{O}_{\boldsymbol{n}} = \{i \in [K] \mid n_i > 0\}$. The number of observed allele types is denoted by $|\mathcal{O}_{\boldsymbol{n}}|$.*

Henceforth, when the indices $h, i, j, k$ and $l$ are used in indefinite summations or products, they are assumed to range over $\mathcal{O}_{\boldsymbol{n}}$, unless stated otherwise.

Given a sample configuration $\boldsymbol{n} = (n_1, \ldots, n_K)$, a $K$-tuple $\boldsymbol{m} = (m_1, \ldots, m_K)$ satisfying $\boldsymbol{0} \preceq \boldsymbol{m} \prec \boldsymbol{n}$ means $0 \leq m_i < n_i$ for all $i \in \mathcal{O}_{\boldsymbol{n}}$ and $m_i = 0$ for all $i \notin \mathcal{O}_{\boldsymbol{n}}$, while $\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}$ means $0 < m_i \leq n_i$ for all $i \in \mathcal{O}_{\boldsymbol{n}}$ and $m_i = 0$ for all $i \notin \mathcal{O}_{\boldsymbol{n}}$. Also, $\boldsymbol{0} \preceq \boldsymbol{m} \preceq \boldsymbol{n}$ denotes $0 \leq m_i \leq n_i$ for all $i \in [K]$.

Rewriting (2.4), we see that $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$ is the unique solution to the recursion

$$n(n-1+\theta)q(\boldsymbol{n} \mid \theta, \boldsymbol{P}) = \sum_i n_i(n_i-1)q(\boldsymbol{n}-\boldsymbol{e}_i \mid \theta, \boldsymbol{P}) + \theta \sum_{i,j} P_{ji}\, n_i\, q(\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j \mid \theta, \boldsymbol{P}), \quad (2.6)$$

with boundary conditions

$$q(\boldsymbol{e}_i \mid \theta, \boldsymbol{P}) = \pi_i, \text{ for all } i \in [K]. \quad (2.7)$$

If $\boldsymbol{P}$ is irreducible when restricted to the observed alleles $\mathcal{O}_{\boldsymbol{n}}$, then by unwinding recursion (2.6), it can be seen that $|\mathcal{O}_{\boldsymbol{n}}| - 1$ is the smallest power of $\theta$ with a non-vanishing coefficient

in the Taylor series expansion of $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$ about $\theta = 0$. Intuitively, for a sample with $m$ distinct observed alleles, the coefficient of $\theta^{m-1}$ in the Taylor expansion corresponds to the total probability of coalescent genealogies with the most parsimonious number (i.e., $m-1$) of mutations. That $\boldsymbol{P}$ is irreducible when restricted to $\mathcal{O}_{\boldsymbol{n}}$ is a sufficient (but not necessary) condition for the existence of such a parsimonious genealogy for sample configuration $\boldsymbol{n}$.

Letting $Q(\boldsymbol{n} \mid \boldsymbol{P})$ denote the coefficient of $\theta^{|\mathcal{O}_{\boldsymbol{n}}|-1}$ in the Taylor expansion, $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$ can be written as in (1.1). For simplicity, in the rest of this thesis, we simply write $q(\boldsymbol{n})$ and $Q(\boldsymbol{n})$ instead of $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$ and $Q(\boldsymbol{n} \mid \boldsymbol{P})$, respectively. In the next chapter, we give explicit formulas for $Q(\boldsymbol{n})$ when at most 4 distinct alleles are observed in the sample $\boldsymbol{n}$, i.e. when $1 \leq |\mathcal{O}_{\boldsymbol{n}}| \leq 4$.

# Chapter 3

# Main results

This chapter summaries our closed-form results for $Q(\boldsymbol{n})$ when at most 4 distinct alleles are observed in the sample, i.e. $|\mathcal{O}_{\boldsymbol{n}}| \leq 4$. In the case of $|\mathcal{O}_{\boldsymbol{n}}| = 1$, it is easy to see that $Q(\boldsymbol{n}) = \pi_i$ for $\boldsymbol{n} = n\boldsymbol{e}_i$. The closed-form results for $Q(\boldsymbol{n})$ when $2 \leq |\mathcal{O}_{\boldsymbol{n}}| \leq 4$ are summarized in the following theorems.

**Theorem 3.** *For $|\mathcal{O}_{\boldsymbol{n}}| = 2$ and $\boldsymbol{P}$ an arbitrary mutation transition matrix that is irreducible when restricted to $\mathcal{O}_{\boldsymbol{n}}$, $Q(\boldsymbol{n})$ is given by*

$$Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n}) \sum_{i,j \in \mathcal{O}_{\boldsymbol{n}}: i \neq j} \frac{n_j}{n} \pi_j P_{ji}.$$

**Theorem 4.** *For $|\mathcal{O}_{\boldsymbol{n}}| = 3$ and $\boldsymbol{P}$ an arbitrary mutation transition matrix that is irreducible when restricted to $\mathcal{O}_{\boldsymbol{n}}$, $Q(\boldsymbol{n})$ is given by*

$$\begin{aligned} Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n}) \sum_{\text{distinct } i,j,k \in \mathcal{O}_{\boldsymbol{n}}} \Bigg\{ & \pi_j P_{ji} P_{jk} \Bigg[ \frac{(n_j)_{2\downarrow}}{n(n_j + n_k - 1)} - \frac{n_i n_j}{n(n_i + n_k)} - 2\frac{n_i n_j n_k}{n(n_j + n_k)_{2\downarrow}} \\ & + 2\frac{n_i n_j n_k}{(n_j + n_k + 1)_{3\downarrow}} (H_n - H_{n_i - 1}) \Bigg] \\ & + \pi_k P_{kj} P_{ji} \Bigg[ \frac{n_j n_k}{n(n_j + n_k - 1)} + 2\frac{n_i n_j n_k}{n(n_j + n_k)_{2\downarrow}} \\ & - 2\frac{n_i n_j n_k}{(n_j + n_k + 1)_{3\downarrow}} (H_n - H_{n_i - 1}) \Bigg] \Bigg\}. \end{aligned}$$

**Corollary 5.** *Suppose $|\mathcal{O}_{\boldsymbol{n}}| = 3$ with sample configuration $\boldsymbol{n} = n_a\boldsymbol{e}_a + n_b\boldsymbol{e}_b + n_c\boldsymbol{e}_c$, where $a, b, c$ are distinct alleles in $[K]$. If the mutation transition matrix $\boldsymbol{P}$ is reversible and irreducible when restricted to the observed alleles $\mathcal{O}_{\boldsymbol{n}}$, $Q(\boldsymbol{n})$ is given by*

$$Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n}) \left( \frac{n_a}{n} \pi_a P_{ab} P_{ac} + \frac{n_b}{n} \pi_b P_{ba} P_{bc} + \frac{n_c}{n} \pi_c P_{ca} P_{cb} \right).$$

**Theorem 6.** *For $|\mathcal{O}_{\boldsymbol{n}}| = 4$, if the mutation transition matrix $\boldsymbol{P}$ is reversible and irreducible when restricted to the observed alleles $\mathcal{O}_{\boldsymbol{n}}$, then $Q(\boldsymbol{n})$ is given by*

$$Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n}) \sum_{\text{distinct } i,j,k,l \in \mathcal{O}_{\boldsymbol{n}}} \left[\pi_i P_{ij} P_{ik} P_{il} \gamma(\boldsymbol{n}, i, j, k, l) + \pi_i P_{ij} P_{ik} P_{jl} \delta(\boldsymbol{n}, i, j, k, l)\right],$$

*where*

$$\begin{aligned}\gamma(\boldsymbol{n}, i, j, k, l) = \frac{n_i}{n} \Bigg\{ &\left[\frac{n_i - 1}{2(n_i + n_j + n_k - 1)} - \frac{2n_j n_l}{(n_i + n_j + n_k)_{2\downarrow}}\right] + \frac{n_l}{2(n_j + n_k + n_l)} \\ &- \left[\frac{n_l(n_i - 1)}{(n_k + n_l)(n_i + n_j - 1)} - \frac{2n_j n_l}{(n_i + n_j)_{2\downarrow}}\right] \Bigg\} \\ &+ \frac{2n_i n_j n_l}{(n_i + n_j + n_k + 1)_{3\downarrow}} (H_n - H_{n_l - 1}) - \frac{2n_i n_j n_l}{(n_i + n_j + 1)_{3\downarrow}} (H_n - H_{n_k + n_l - 1}),\end{aligned}$$

*and*

$$\begin{aligned}\delta(\boldsymbol{n}, i, j, k, l) = \frac{n_i}{n} \Bigg\{ &\left[\frac{n_j}{n_i + n_j + n_k - 1} + \frac{2n_j n_l}{(n_i + n_j + n_k)_{2\downarrow}}\right] \\ &- \left[\frac{n_j n_l}{(n_k + n_l)(n_i + n_j - 1)} + \frac{2n_j n_l}{(n_i + n_j)_{2\downarrow}}\right] \Bigg\} \\ &- \frac{2n_i n_j n_l}{(n_i + n_j + n_k + 1)_{3\downarrow}} (H_n - H_{n_l - 1}) + \frac{2n_i n_j n_l}{(n_i + n_j + 1)_{3\downarrow}} (H_n - H_{n_k + n_l - 1}).\end{aligned}$$

Note that by precomputing $H_k$ and $k!$ (or more suitably, $\log k!$) for $1 \le k \le n$, all the formulas in the above theorems can be evaluated essentially in constant time and space for any sample of size at most $n$.

By restricting the set of events in the coalescent genealogy for a given sample, Jenkins and Song [21] have also provided closed-form formulas for $Q(\boldsymbol{n} \mid \boldsymbol{P})$ for an arbitrary transition matrix $\boldsymbol{P}$ when $|\mathcal{O}_{\boldsymbol{n}}| \le 3$. We provide new proofs of those results, while Theorem 6 extends them by giving a closed-form formula for $Q(\boldsymbol{n} \mid \boldsymbol{P})$ when $|\mathcal{O}_{\boldsymbol{n}}| = 4$. As there are four distinct DNA bases, our extension to the $|\mathcal{O}_{\boldsymbol{n}}| = 4$ case seems natural.

Jenkins and Song [21] have also shown that a simple generalization of the formulas in Theorem 3 and Corollary 5 hold for arbitrary values of $|\mathcal{O}_{\boldsymbol{n}}|$ when $\boldsymbol{P}$ is parent-independent restricted to the observed alleles in $\boldsymbol{n}$. More precisely, for mutation matrices $\boldsymbol{P}$ where $P_{ji} = P_{ki}$ for all $i, j, k \in \mathcal{O}_{\boldsymbol{n}}$, they showed that $Q(\boldsymbol{n})$ is given by,

$$Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n}) \sum_i \left(\frac{n_i}{n} \pi_i \prod_{j \ne i} P_{ij}\right). \tag{3.1}$$

From Corollary 5, it can be seen that the formula (3.1) also holds when $\boldsymbol{P}$ is reversible restricted to the observed alleles, provided that $|\mathcal{O}_{\boldsymbol{n}}| \le 3$. However, that formula fails to hold when $|\mathcal{O}_{\boldsymbol{n}}| = 4$ and $\boldsymbol{P}$ is reversible but not parent-independent restricted to the observed alleles, as can be seen from Theorem 6.

# Chapter 4

# Proofs of the main results

To prove the closed-form formulas for $Q(\boldsymbol{n})$ given in Chapter 3, we will construct an urn process that is intimately related to the coalescent. We use this urn process to decompose $Q(\boldsymbol{n})$ into a sum-product of two vectors, one which depends only on the sample configuration $\boldsymbol{n}$ and the other which depends only on the mutation transition matrix $\boldsymbol{P}$. Using this decomposition, we show that $Q(\boldsymbol{n})$ corresponds to the probability of a certain event in the urn process. Throughout, we use $R(\boldsymbol{n})$ to denote the following rescaled version of $Q(n)$:

$$R(\boldsymbol{n}) = \frac{Q(\boldsymbol{n})}{\Lambda(\boldsymbol{n})}, \tag{4.1}$$

where $\Lambda(\boldsymbol{n})$ is the combinatorial coefficient defined in (2.1).

## 4.1 An urn construction

Let $\boldsymbol{n}$ be the sample configuration of interest. We have an urn with $n$ balls, $n_i$ of which have color $i$. We remove balls one at a time uniformly at random until there are no more balls in the urn. However, whenever we "kill" a color (i.e., remove the last ball of that color), we add back a ball of a different color. We do this by picking another ball from the urn, copying it, and returning both copies to the urn. Note that when we kill the last color, we do not add any balls back, since there are no more colors to choose from.

Suppose that when we kill color $i$, we add back a ball of color $j$. We then call $j$ the *parent* of $i$, and call the last surviving color the *root*. We abuse terminology and also call the particular ball of color $j$ as the parent of color $i$. This process generates a rooted tree whose vertices consist of the $|\mathcal{O}_{\boldsymbol{n}}|$ observed colors (alleles).

Let $T$ be any rooted tree on $\mathcal{O}_{\boldsymbol{n}}$. We denote the probability of generating $T$ under the above process as $\mathbb{P}_{\boldsymbol{n}}(T)$. Let $E(T)$ be the edge set of $T$, and let $\rho(T)$ denote the root vertex of $T$. By convention, we draw edges as pointing away from the root, so the edge $(j \to i)$ indicates that $j$ is the parent of $i$.

The main idea of this section is that to compute $Q(\boldsymbol{n})$, it is enough to compute $\mathbb{P}_{\boldsymbol{n}}(T)$ for each $T$. In particular, we prove the following theorem in Section 4.1:

**Theorem 7.** *Recall that for a transition matrix $\boldsymbol{P}$ that is irreducible when restricted to $\mathcal{O}_{\boldsymbol{n}}$, $Q(\boldsymbol{n})$ denotes the first nonzero coefficient in the Taylor expansion (1.1) of $q(\boldsymbol{n})$ about $\theta = 0$. Given a rooted tree $T$ described above, define $f_{\boldsymbol{P}}(T)$ as*

$$f_{\boldsymbol{P}}(T) = \pi_{\rho(T)} \prod_{(j \to i) \in E(T)} P_{ji}.$$

*Then, the quantity $R(\boldsymbol{n}) = Q(\boldsymbol{n})/\Lambda(\boldsymbol{n})$ is given by*

$$R(\boldsymbol{n}) = \sum_T \mathbb{P}_{\boldsymbol{n}}(T) f_{\boldsymbol{P}}(T) = \mathbb{E}_{\boldsymbol{n}}[f_{\boldsymbol{P}}(T)], \tag{4.2}$$

*where the sum is taken over all rooted trees $T$ with $|\mathcal{O}_{\boldsymbol{n}}|$ vertices bijectively labeled by $\mathcal{O}_{\boldsymbol{n}}$. That is, $R(\boldsymbol{n})$ is the expectation of $f_{\boldsymbol{P}}(T)$ under the above process.*

Note that we can view $f_{\boldsymbol{P}}(T)$ as a probability as well. In particular, suppose we relabel the vertices of $T$ as follows: we assign a new label from $[K]$ to $\rho(T)$ according to the stationary distribution $\pi$, and for each edge in $T$, we assign a new label to the child according to the new label of its parent and the transition matrix $\boldsymbol{P}$. Then $f_{\boldsymbol{P}}(T)$ is the probability that we assign the original labels to all the vertices, given that we drew $T$. That is, if $\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}}$ is the event that we assign the original labels to all vertices, then

$$f_{\boldsymbol{P}}(T) = \mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T) = \pi_{\rho(T)} \prod_{(j \to i) \in E(T)} P_{ji}.$$

This immediately leads to the following interpretation:

$$R(\boldsymbol{n}) = \sum_T \mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T) \mathbb{P}_{\boldsymbol{n}}(T) = \mathbb{P}_{\boldsymbol{n}}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}}). \tag{4.3}$$

That is, $R(\boldsymbol{n})$ is the unconditional probability that we correctly label all the alleles, if we use the urn process to generate a tree on the alleles and then use the tree to assign labels.

## An inductive proof of Theorem 7

In this subsection, we provide an inductive proof of Theorem 7. In Section 4.2, we provide an alternative proof based on a modified coalescent process which provides a more intuitive explanation for why the urn process works.

*Proof of Theorem 7.* Recall the recursion in (2.6):

$$n(n-1+\theta) q(\boldsymbol{n}) = \sum_i n_i (n_i - 1) q(\boldsymbol{n} - \boldsymbol{e}_i) + \theta \sum_{i,j} P_{ji}\, n_i\, q(\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j).$$

Recall also that if $\boldsymbol{P}$ is irreducible when restricted to $\mathcal{O}_{\boldsymbol{n}}$, $q(\boldsymbol{n})$ has leading order power $\theta^{|\mathcal{O}_{\boldsymbol{n}}|-1}$ in its Taylor series. Hence we get the following recursion for $Q(\boldsymbol{n})$:

$$n(n-1)Q(\boldsymbol{n}) = \sum_{i:n_i>1} n_i(n_i-1)Q(\boldsymbol{n}-\boldsymbol{e}_i) + \sum_{i:n_i=1}\sum_{j:j\neq i} P_{ji}n_iQ(\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j).$$

Plugging in $Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n})R(\boldsymbol{n})$ and simplifying gives us the following recursion for $R(\boldsymbol{n})$:

$$n(n-1)R(\boldsymbol{n}) = \sum_{i:n_i>1} n_i(n-1)R(\boldsymbol{n}-\boldsymbol{e}_i) + \sum_{i:n_i=1}\sum_{j:j\neq i} P_{ji}n_jR(\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j). \tag{4.4}$$

A simple induction over $|\mathcal{O}_{\boldsymbol{n}}|$ and $n$ shows that this recursion has a unique solution given the boundary conditions $R(\boldsymbol{e}_i)$. So if we can show (4.2) when $|\mathcal{O}_{\boldsymbol{n}}| = n = 1$, and then show that $\sum_T \mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T)\mathbb{P}_{\boldsymbol{n}}(T)$ satisfies the recursion (4.4), then we will be done. The base case is trivial: when $\mathcal{O}_{\boldsymbol{n}} = \{a\}$, there is only one possible tree, $T = \{a\}$, with $\mathbb{P}_{\boldsymbol{n}}(T) = 1$ and $\mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T) = \pi_a = \lim_{\theta\to 0} q(\boldsymbol{n}) = Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n})R(\boldsymbol{n}) = R(\boldsymbol{n})$.

To show $\sum_T \mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T)\mathbb{P}_{\boldsymbol{n}}(T)$ satisfies (4.4), we start by giving recursions for $\mathbb{P}_{\boldsymbol{n}}(T)$ and $\mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T)$. Let $z(i)$ be the parent of $i$ in $T$, and let $L(T)$ be the set of leafs of $T$ (where the root is not considered a leaf). Conditioning on the first event in the urn process gives us

$$\mathbb{P}_{\boldsymbol{n}}(T) = \sum_{i:n_i>1} \frac{n_i}{n}\mathbb{P}_{\boldsymbol{n}-\boldsymbol{e}_i}(T) + \sum_{i\in L(T):n_i=1} \frac{n_{z(i)}}{n(n-1)}\mathbb{P}_{\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_{z(i)}}(T\setminus\{i\}). \tag{4.5}$$

Furthermore, if $i \in L(T)$, we have

$$\mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T) = P_{z(i),i}\mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}\setminus\{i\}} \mid T\setminus\{i\}). \tag{4.6}$$

Using (4.5) and (4.6), and collecting terms, we arrive at

$$\begin{aligned}
&n(n-1)\sum_T \mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T)\mathbb{P}_{\boldsymbol{n}}(T)\\
&\quad= \sum_T \mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T)\left[\sum_{i:n_i>1} n_i(n-1)\mathbb{P}_{\boldsymbol{n}-\boldsymbol{e}_i}(T) + \sum_{i\in L(T):n_i=1} n_{z(i)}\mathbb{P}_{\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_{z(i)}}(T\setminus\{i\})\right]\\
&\quad= \sum_{i:n_i>1} n_i(n-1)\sum_T \mathbb{P}_{\boldsymbol{n}-\boldsymbol{e}_i}(T)\mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T)\\
&\qquad+ \sum_{i:n_i=1}\sum_{j:j\neq i} P_{ji}n_j \sum_{T'} \mathbb{P}_{\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j}(T')\mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}\setminus\{i\}} \mid T'),
\end{aligned}$$

where the sum over $T'$ is taken over all rooted trees with vertex set $\mathcal{O}_{\boldsymbol{n}} \setminus \{i\}$. Hence, $\sum_T \mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T)\mathbb{P}_{\boldsymbol{n}}(T)$ satisfies (4.4). □

## 4.2 Connection to the coalescent

In this section, we motivate our urn process by drawing a connection to the coalescent. We then use this connection with the coalescent to provide an alternate proof of Theorem 7.

Let $\mathcal{H}$ be a history of mutation and coalescence events on $n$ labeled individuals, and let $q(\mathcal{H})$ be the probability of $\mathcal{H}$. Then we have

$$q(\boldsymbol{n}) = \sum_{\mathcal{H} \text{ consistent with } \boldsymbol{n}} q(\mathcal{H}). \tag{4.7}$$

It turns out that only histories with exactly $|\mathcal{O}_{\boldsymbol{n}}| - 1$ mutations contribute to the leading order term of $q(\boldsymbol{n})$; this is the observation also utilized in [21]. Furthermore, each history of choices in our urn process corresponds with a genealogical history of $|\mathcal{O}_{\boldsymbol{n}}| - 1$ mutations. This provides the basic intuition for why the urn sampling scheme works.

We start by providing a modified coalescent that generates a history $\mathcal{H}$ that is consistent with $\boldsymbol{n}$ and has exactly $|\mathcal{O}_{\boldsymbol{n}}| - 1$ mutations. We then show that this modified coalescent is equivalent to our urn sampling process. Finally, we prove Theorem 7 by relating the modified coalescent with Kingman's coalescent.

Consider the following modified coalescent process on our sample:

1. Select allele $i$ with probability $m_i/m$, where $\boldsymbol{m}$ is our current configuration of alleles.
2. If $m_i > 1$, choose a random pair in allele $i$ to coalesce (so $\boldsymbol{m}$ is replaced with $\boldsymbol{m} - \boldsymbol{e}_i$).
3. If $m_i = 1$, have the last individual of allele $i$ mutate to allele $j$ with probability $m_j/(m-1)$ (so $\boldsymbol{m}$ is replaced with $\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j$).
4. Repeat steps 1 to 3 until all individuals have coalesced.

It should be clear that the modified coalescent only generates histories with exactly $|\mathcal{O}_{\boldsymbol{n}}| - 1$ mutations, since each mutation kills an allele permanently.

If we take an unordered view of our sample, then the modified coalescent is equivalent to the urn process, for they have the same initial configuration and transition probabilities between configurations. In particular, when $m_i > 1$ we move from $\boldsymbol{m}$ to $\boldsymbol{m} - \boldsymbol{e}_i$ with probability $m_i/m$, and when $m_i = 1$ we move from $\boldsymbol{m}$ to $\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j$ with probability $m_j/(m)_{2\downarrow}$. We generate trees on $\mathcal{O}_{\boldsymbol{n}}$ by drawing an edge $(j \to i)$ whenever we make a transition from $\boldsymbol{m}$ to $\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j$, i.e. whenever there is a mutation from $i$ to $j$.

We now give a proof of Theorem 7, using the modified coalescent in place of the urn process:

*Alternative proof of Theorem 7.* Let $\mathcal{H}$ be a coalescent history with exactly $M$ mutations. Running time backwards from the present, we suppose that the $i$th mutation was from allele $u_i$ to allele $v_i$, and that the most recent common ancestor has allele $\rho$. We further suppose

that $J_i$ is the total number of lineages at the time of the $i$th mutation. Then we have that

$$q(\mathcal{H}) = \pi_\rho\Big(\prod_{i=1}^{M} P_{v_i u_i}\Big)\frac{\theta^M}{\prod_{i=1}^{M} J_i(\theta + J_i - 1)}\frac{2^{n-1}}{n!(\theta+n-1)_{(n-1)\downarrow}},$$

since the $i$th coalescence contributes probability $\frac{n-i}{n-i+\theta}\binom{n-i+1}{2}^{-1} = \frac{2}{(n-i+1)(n-i+\theta)}$, and the $i$th mutation contributes probability $\frac{\theta P_{v_i u_i}}{J_i(J_i-1+\theta)}$.

Now, observe that

$$Q(\mathcal{H}) \equiv \lim_{\theta\to 0}\frac{q(\mathcal{H})}{\theta^M} = \pi_\rho\Big(\prod_{i=1}^{M} P_{v_i u_i}\Big)\frac{2^{n-1}}{n!(n-1)!\prod_{i=1}^{M} J_i(J_i-1)}. \tag{4.8}$$

Therefore, the Taylor series for $q(\mathcal{H})$ has leading power $\theta^M$, with coefficient $Q(\mathcal{H})$.

Hence by (4.7), the Taylor series for $q(\boldsymbol{n})$ has leading power $\theta^{|\mathcal{O}_{\boldsymbol{n}}|-1}$, and its leading coefficient is given by the sum of all $Q(\mathcal{H})$ such that $\mathcal{H}$ is consistent with $\boldsymbol{n}$ and has $|\mathcal{O}_{\boldsymbol{n}}|-1$ mutations.

For such an $\mathcal{H}$, let $\mathbb{P}_{\boldsymbol{n}}(\mathcal{H})$ be the probability of generating $\mathcal{H}$ under our modified coalescent. Then we have that

$$\mathbb{P}_{\boldsymbol{n}}(\mathcal{H}) = \frac{2^{n-1}}{n!\prod_{k=1}^{|\mathcal{O}_{\boldsymbol{n}}|}(n_k-1)!\prod_{i=1}^{|\mathcal{O}_{\boldsymbol{n}}|-1} J_i(J_i-1)}. \tag{4.9}$$

To see this, note that if our current sample is $\boldsymbol{m}$, the probability that the next event is a coalescence on allele $i$ with $m_i > 1$ is

$$\frac{m_i}{m}\frac{2}{m_i(m_i-1)} = \frac{2}{m(m_i-1)},$$

and if $m_i = 1$, the probability that the next event is a mutation from allele $i$ to allele $j$ (where $j \neq i$) is

$$\frac{m_j}{m(m-1)}.$$

Multiplying the probabilities of the mutation and coalescence events in $\mathcal{H}$, and noting that the numerator of each mutation term cancels with the denominator of a future coalescence term, yields the equation (4.9).

Combining (4.8) with (4.9) yields

$$Q(\mathcal{H}) = \Lambda(\boldsymbol{n})\pi_\rho\Big(\prod_{i=1}^{|\mathcal{O}_{\boldsymbol{n}}|-1} P_{v_i u_i}\Big)\mathbb{P}_{\boldsymbol{n}}(\mathcal{H})$$

Now let $\mathscr{T}(\mathcal{H})$ be the resulting tree on $\mathcal{O}_{\boldsymbol{n}}$ if we draw an edge $(j \to i)$ when allele $i$ mutates to allele $j$. Then we have

$$\begin{aligned} Q(\boldsymbol{n}) &= \sum_{\substack{\mathcal{H} \text{ consistent with } \boldsymbol{n} \\ \mathcal{H} \text{ has } |\mathcal{O}_{\boldsymbol{n}}|-1 \text{ mutations}}} Q(\mathcal{H}) \\ &= \Lambda(\boldsymbol{n}) \sum_T \pi_{\rho(T)} \Big( \prod_{(j\to i)\in T} P_{ji} \Big) \Big( \sum_{\mathcal{H}:\mathscr{T}(\mathcal{H})=T} \mathbb{P}_{\boldsymbol{n}}(\mathcal{H}) \Big) \\ &= \Lambda(\boldsymbol{n}) \sum_T \pi_{\rho(T)} \Big( \prod_{(j\to i)\in T} P_{ji} \Big) \mathbb{P}_{\boldsymbol{n}}(T) \\ &= \Lambda(\boldsymbol{n}) \sum_T f_{\boldsymbol{P}}(T) \mathbb{P}_{\boldsymbol{n}}(T), \end{aligned}$$

and hence

$$R(\boldsymbol{n}) = \sum_T f_{\boldsymbol{P}}(T) \mathbb{P}_{\boldsymbol{n}}(T),$$

as needed. □

## 4.3 Martingale proofs of Theorem 3 and Corollary 5

Here, we prove Theorem 3 and Corollary 5 by using a martingale argument to compute $\mathbb{P}_{\boldsymbol{n}}(T)$ for $\mathcal{O}_{\boldsymbol{n}} = \{a, b\}$, and for $\mathcal{O}_{\boldsymbol{n}} = \{a, b, c\}$ when $\boldsymbol{P}$ is reversible restricted to $\mathcal{O}_{\boldsymbol{n}}$. We run time as follows: whenever we remove a ball in the urn process, we count this as one time step. If in doing so, we kill a color, we count the adding of another ball as a separate time step.

Let $\mathcal{F}_t$ be the $\sigma$-algebra generated by all sequences of choices up to time $t$. Let $X_t$ be the proportion of balls that have color $a$ at time $t$; so $X_0 = n_a/n$. It is easy to check that $\{X_t\}$ is a martingale with respect to $\{\mathcal{F}_t\}$: suppose that $\boldsymbol{m}$ is the remaining sample after time $t-1$, and we are deleting a ball at time $t$. Then,

$$\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \frac{m_a}{m}\frac{m_a - 1}{m-1} + \sum_{i\neq a} \frac{m_i}{m}\frac{m_a}{m-1} = \frac{m_a}{m} = X_{t-1}.$$

On the other hand, if we are adding a ball at time $t$, then

$$\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \frac{m_a}{m}\frac{m_a + 1}{m+1} + \sum_{i\neq a} \frac{m_i}{m}\frac{m_a}{m+1} = \frac{m_a}{m} = X_{t-1}.$$

So, $\{(X_t, \mathcal{F}_t), t \geq 0\}$ is a martingale.

*Proof of Theorem 3.* Suppose $\mathcal{O}_{\boldsymbol{n}} = \{a, b\}$. Let $T$ be the tree whose vertex set is $\mathcal{O}_{\boldsymbol{n}}$, with $a$ being the root. Let $\tau$ be the the first time we kill a color. Noting that $\tau$ is a stopping time,

we obtain

$$\begin{aligned}\mathbb{P}_{\boldsymbol{n}}(T) &= \mathbb{E}[\mathbb{P}_{\boldsymbol{n}}(T \mid \mathcal{F}_\tau)] \\ &= \mathbb{E}[\mathbb{I}(\text{Color } a \text{ is the last remaining at time } \tau)] \\ &= \mathbb{E}[X_\tau] = \mathbb{E}[X_0] = \frac{n_a}{n}.\end{aligned}$$

Therefore, by Theorem 7,

$$Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n})\Big(\frac{n_a}{n}\pi_a P_{ab} + \frac{n_b}{n}\pi_b P_{ba}\Big). \qquad \square$$

*Proof of Corollary 5.* Suppose $\mathcal{O}_{\boldsymbol{n}} = \{a, b, c\}$ and $\boldsymbol{P}$ is reversible when restricted to $\mathcal{O}_{\boldsymbol{n}}$. Note that $\mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T)$ does not depend on how $T$ is rooted, for by reversibility we can move the root around by

$$\pi_\rho P_{\rho k} = \pi_k P_{k\rho}, \qquad \forall k \in \mathcal{O}_{\boldsymbol{n}}, k \neq \rho.$$

Therefore, we redefine $\mathbb{P}_{\boldsymbol{n}}(T)$ to be the probability of drawing the undirected tree $T$. We still have $R(\boldsymbol{n}) = \sum_T \mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid T)\mathbb{P}_{\boldsymbol{n}}(T)$, but now the sum is taken over undirected $T$. Now let $T$ be the tree on $\{a, b, c\}$ whose interior vertex is $a$. We draw $T$ if and only if $a$ is chosen as the parent of the first color that we kill. So, letting $\tau$ be the first killing time and noting $X_\tau = \mathbb{P}_{\boldsymbol{n}}(T \mid \mathcal{F}_\tau)$, we have

$$\mathbb{P}_{\boldsymbol{n}}(T) = \mathbb{E}[\mathbb{P}_{\boldsymbol{n}}(T \mid \mathcal{F}_\tau)] = \mathbb{E}[X_\tau] = \mathbb{E}[X_0] = \frac{n_a}{n}.$$

Therefore, by Theorem 7,

$$Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n})\Big(\frac{n_a}{n}\pi_a P_{ab}P_{ac} + \frac{n_b}{n}\pi_b P_{ba}P_{bc} + \frac{n_c}{n}\pi_c P_{ca}P_{cb}\Big). \qquad \square$$

## 4.4 Recursion for the approximate sampling probability

In this section, we derive a recursion for $R(\boldsymbol{n})$ which will be useful for deriving closed-form formulas for $Q(\boldsymbol{n})$ when $|\mathcal{O}_{\boldsymbol{n}}| = 3, 4$. Given a sample configuration $\boldsymbol{n}$ and a subsample $\boldsymbol{m}$, define the expression $\binom{\boldsymbol{n}}{\boldsymbol{m}}$ as

$$\binom{\boldsymbol{n}}{\boldsymbol{m}} = \prod_{i \in \mathcal{O}_{\boldsymbol{n}}} \binom{n_i}{m_i}.$$

The following proposition provides a recursion relating $R(\boldsymbol{n})$ to $R(\boldsymbol{m})$ where $|\mathcal{O}_{\boldsymbol{m}}| = |\mathcal{O}_{\boldsymbol{n}}| - 1$.

**Proposition 8.** *Suppose $\boldsymbol{P}$ is irreducible when restricted to $\mathcal{O}_{\boldsymbol{n}}$ and let $\theta^{|\mathcal{O}_{\boldsymbol{n}}|-1}Q(\boldsymbol{n}) = \theta^{|\mathcal{O}_{\boldsymbol{n}}|-1}\Lambda(\boldsymbol{n})R(\boldsymbol{n})$ denote the leading order term in the Taylor expansion (1.1) of $q(\boldsymbol{n})$ about*

$\theta = 0$. *Then,* $R(\boldsymbol{n})$ *for* $|\mathcal{O}_{\boldsymbol{n}}| > 1$ *satisfies the recursion*

$$R(\boldsymbol{n}) = \sum_{i,j \in \mathcal{O}_{\boldsymbol{n}}: i \neq j} P_{ji} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_j R(\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j)}{m(m-1)}, \tag{4.10}$$

*with boundary conditions*

$$R(\boldsymbol{n}) = \pi_a, \tag{4.11}$$

*for all sample configurations* $\boldsymbol{n} = n_a \boldsymbol{e}_a$*, where* $a \in [K]$.

*Proof of Proposition 8.* We can derive this recursion from the urn process as follows. Let $D_{ij}(\boldsymbol{m})$ be the event where the first killing replaces a ball of color $i$ with a ball of color $j$, and where $\boldsymbol{m}$ is the (unordered) configuration immediately before this killing. Then for any event $A$,

$$\mathbb{P}_{\boldsymbol{n}}(A) = \sum_{i,j \neq i} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1}} \mathbb{P}_{\boldsymbol{n}}(D_{ij}(\boldsymbol{m})) \mathbb{P}_{\boldsymbol{n}}(A \mid D_{ij}(\boldsymbol{m})) \tag{4.12}$$

where we use the fact that $\mathbb{P}_{\boldsymbol{n}}(D_{ij}(\boldsymbol{m})) = 0$ if $m_i \neq 1$ or $m_j = 0$ for any $j \in \mathcal{O}_{\boldsymbol{n}}$.

We compute $\mathbb{P}_{\boldsymbol{n}}(D_{ij}(\boldsymbol{m}))$ when $\boldsymbol{m} \succ \boldsymbol{0}$ and $m_i = 1$. The probability that $\boldsymbol{m}$ is the remaining configuration after $n - m$ draws is

$$\frac{(n-m)!}{\prod_k (n_k - m_k)!} \frac{\prod_k (n_k)_{n_k - m_k\downarrow}}{(n)_{n-m\downarrow}} = \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}}.$$

To see this, note that the first term is the number of ways we can make $n - m$ draws that result in the configuration $\boldsymbol{m}$, and the second term is the probability of each such sequence of draws.

When our current configuration is $\boldsymbol{m}$ with $m_i = 1$, the probability that on the next draw we replace the last ball of color $i$ with a ball of color $j$ is $m_j/(m)_{2\downarrow}$. Hence we get that

$$\mathbb{P}_{\boldsymbol{n}}(D_{ij}(\boldsymbol{m})) = \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_j}{m(m-1)}.$$

when $\boldsymbol{m} \succ \boldsymbol{0}$ and $m_i = 1$.

Plugging this into (4.12) yields

$$\mathbb{P}_{\boldsymbol{n}}(A) = \sum_{i,j \neq i} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_j}{m(m-1)} \mathbb{P}_{\boldsymbol{n}}(A \mid D_{ij}(\boldsymbol{m})). \tag{4.13}$$

Now recall from (4.3) that $R(\boldsymbol{n}) = \mathbb{P}_{\boldsymbol{n}}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}})$. That is, $R(\boldsymbol{n})$ is the probability that we assign the original labels to all alleles, if we use the urn process to generate a tree on $\mathcal{O}_{\boldsymbol{n}}$ and then use the tree to assign new labels to the alleles. Note that

$$\mathbb{P}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}} \mid D_{ij}(\boldsymbol{m})) = P_{ji} \mathbb{P}_{\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j}(\mathcal{C}_{\mathcal{O}_{\boldsymbol{n}} \setminus \{i\}}) = P_{ji} R(\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j),$$

since we need to use the urn process with sample $\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j$ to correctly relabel $\mathcal{O}_{\boldsymbol{n}} \setminus \{i\}$, and then assign the correct label to $\{i\}$ with probability $P_{ji}$. Plugging this into (4.13) with $A = \mathcal{C}_{\mathcal{O}_{\boldsymbol{n}}}$ yields the desired recursion,

$$R(\boldsymbol{n}) = \sum_{i,j \neq i} P_{ji} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_j R(\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j)}{m(m-1)}. \qquad \square$$

In Section 4.5 and Section 4.7, we use the recursion in Proposition 8 to provide proofs of Theorem 4 and Theorem 6.

## 4.5 Proof of Theorem 4 ($|\mathcal{O}_{\boldsymbol{n}}| = 3$)

For $|\mathcal{O}_{\boldsymbol{n}}| = 3$, the following expression for $R(\boldsymbol{n})$ can be derived using Proposition 8:

$$\begin{aligned}
R(\boldsymbol{n}) &= \sum_{i,j \neq i} P_{ji} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_j R(\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j)}{m(m-1)} \\
&= \sum_{i,j \neq i} P_{ji} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_j}{m(m-1)} \sum_{\substack{k,l: \\ l \neq k \text{ and } k,l \neq i}} \frac{m_k + \delta_{j,k}}{m} \pi_k P_{kl} \\
&= \sum_{i,j \neq i} P_{ji} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1}} \left\{ \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{1}{m^2(m-1)} \right. \\
&\qquad \left. \times \left[ \sum_{l: l \neq i,j} m_j(m_j+1) \pi_j P_{jl} + \sum_{k: k \neq i,j} m_j m_k \pi_k P_{kj} \right] \right\} \\
&= \sum_{i,j \neq i} P_{ji} \sum_{m=3}^{n} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1, |\boldsymbol{m}| = m}} \left\{ \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{1}{m^2(m-1)} \right. \\
&\qquad \left. \times \left[ \sum_{k: k \neq i,j} m_j(m_j+1) \pi_j P_{jk} + \sum_{k: k \neq i,j} m_j m_k \pi_k P_{kj} \right] \right\} \\
&= \sum_{i,j,k \text{ distinct}} \sum_{m=3}^{n} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1, |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{\pi_j P_{ji} P_{jk} m_j(m_j+1) + \pi_k P_{kj} P_{ji} m_j m_k}{m^2(m-1)}, \qquad (4.14)
\end{aligned}$$

where in the second equality, the formula from Theorem 3 is used, noting that $|\mathcal{O}_{\boldsymbol{m} - \boldsymbol{e}_i + \boldsymbol{e}_j}| = 2$. If we define the quantities $\alpha(\boldsymbol{n}, i, j, k)$ and $\beta(\boldsymbol{n}, i, j, k)$ as

$$\alpha(\boldsymbol{n}, i, j, k) = \sum_{m=3}^{n} \frac{1}{m^2(m-1)} \sum_{\substack{\boldsymbol{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1, |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} m_j(m_j+1), \qquad (4.15)$$

and

$$\beta(\boldsymbol{n},i,j,k)=\sum_{m=3}^{n}\frac{1}{m^2(m-1)}\sum_{\substack{\mathbf{0}\prec\boldsymbol{m}\preceq\boldsymbol{n}:\\ m_i=1,|\boldsymbol{m}|=m}}\frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}}m_jm_k,$$

then (4.14) can be rewritten as

$$R(\boldsymbol{n})=\sum_{i,j,k\text{ distinct}}\pi_jP_{ji}P_{jk}\alpha(\boldsymbol{n},i,j,k)+\sum_{i,j,k\text{ distinct}}\pi_kP_{kj}P_{ji}\beta(\boldsymbol{n},i,j,k). \tag{4.16}$$

Now consider $\alpha(\boldsymbol{n},i,j,k)$ defined by (4.15). We can remove the restriction in the inner sum that $m_i=1$ by defining $\boldsymbol{m}'=\boldsymbol{m}-\boldsymbol{e}_i$, and so $|\boldsymbol{m}'|=m-1$. Also, since $j\neq i$ in (4.16), $m'_j=m_j$. Making this change of variables from $\boldsymbol{m}$ to $\boldsymbol{m}'$ in the inner sum of (4.15), we get

$$\sum_{\substack{\mathbf{0}\prec\boldsymbol{m}\preceq\boldsymbol{n}:\\ m_i=1,|\boldsymbol{m}|=m}}\frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}}m_j(m_j+1)=\frac{\binom{n-n_i}{m-1}}{\binom{n}{m}}n_i\sum_{\substack{\mathbf{0}\prec\boldsymbol{m}'\preceq\boldsymbol{n}-n_i\boldsymbol{e}_i:\\ |\boldsymbol{m}'|=m-1}}\frac{\binom{\boldsymbol{n}-n_i\boldsymbol{e}_i}{\boldsymbol{m}'}}{\binom{n-n_i}{m-1}}m'_j(m'_j+1). \tag{4.17}$$

Using identity (A.4) in Fact 5 of the Appendix, the summation over $\boldsymbol{m}'$ in (4.17) can be written as

$$\begin{aligned}&\sum_{\substack{\mathbf{0}\prec\boldsymbol{m}'\preceq\boldsymbol{n}-n_i\boldsymbol{e}_i:\\ |\boldsymbol{m}'|=m-1}}\frac{\binom{\boldsymbol{n}-n_i\boldsymbol{e}_i}{\boldsymbol{m}'}}{\binom{n-n_i}{m-1}}m'_j(m'_j+1)\\ &\qquad=\sum_{\substack{T\subseteq[L]:\\ i,j\notin T}}(-1)^{|T|}\left[\frac{(n_j)_{2\downarrow}(m-1)_{2\downarrow}}{(n-n_i-n_T)_{2\downarrow}}+\frac{2n_j(m-1)}{n-n_i-n_T}\right]\frac{\binom{n-n_i-n_T}{m-1}}{\binom{n-n_i}{m-1}}\end{aligned} \tag{4.18}$$

The only sets $T$ satisfying the conditions in the summation in (4.18) are $T = \varnothing$ and $T = \{k\}$. Hence, substituting (4.17) and (4.18) in (4.15), we have

$$\begin{aligned}
\alpha(\boldsymbol{n}, i, j, k) &= \sum_{m=3}^{n} \frac{1}{m^2(m-1)} \sum_{\substack{\mathbf{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1, |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} m_j(m_j+1) \\
&= \sum_{m=3}^{n} \frac{1}{m^2(m-1)} \frac{\binom{n-n_i}{m-1}}{\binom{n}{m}} n_i \sum_{\substack{\mathbf{0} \prec \boldsymbol{m}' \preceq \boldsymbol{n}-n_i\boldsymbol{e}_i: \\ |\boldsymbol{m}'| = m-1}} \frac{\binom{\boldsymbol{n}-n_i\boldsymbol{e}_i}{\boldsymbol{m}'}}{\binom{n-n_i}{m-1}} m'_j(m'_j+1) \\
&= \sum_{m=3}^{n} \frac{n_i \binom{n-n_i}{m-1}}{m^2(m-1)\binom{n}{m}} \left\{ \frac{\binom{n_j+n_k}{m-1}}{\binom{n-n_i}{m-1}} \left[ \frac{(n_j)_{2\downarrow}}{(n_j+n_k)_{2\downarrow}} (m-1)_{2\downarrow} + 2\frac{n_j(m-1)}{n_j+n_k} \right] \right. \\
&\qquad \left. - \frac{\binom{n_j}{m-1}}{\binom{n-n_i}{m-1}} \left[ \frac{(n_j)_{2\downarrow}}{(n_j)_{2\downarrow}} (m-1)_{2\downarrow} + 2\frac{n_j}{n_j}(m-1) \right] \right\} \\
&= \sum_{m=3}^{n} \frac{n_i}{m^2(m-1)} \left\{ \frac{\binom{n_j+n_k}{m-1}}{\binom{n}{m}} \left[ \frac{(n_j)_{2\downarrow}}{(n_j+n_k)_{2\downarrow}} (m-1)_{2\downarrow} + 2\frac{n_j(m-1)}{n_j+n_k} \right] \right. \\
&\qquad \left. - \frac{\binom{n_j}{m-1}}{\binom{n}{m}} m(m-1) \right\} \\
&= \sum_{m=1}^{n} \frac{n_i}{n} \left\{ \frac{\binom{n_j+n_k}{m}}{\binom{n-1}{m}} \left[ \frac{(n_j)_{2\downarrow}}{(n_j+n_k)_{2\downarrow}} \frac{m-1}{m+1} + 2\frac{n_j}{n_j+n_k} \frac{1}{m+1} \right] - \frac{\binom{n_j}{m}}{\binom{n-1}{m}} \right\}. \qquad (4.19)
\end{aligned}$$

Applying Facts 1 and 3 in the Appendix to (4.19) yields

$$\begin{aligned}
\alpha(\boldsymbol{n}, i, j, k) &= \sum_{m=1}^{n} \frac{n_i}{n} \left[ \frac{(n_j)_{2\downarrow}}{(n_j+n_k)_{2\downarrow}} \frac{\binom{n_j+n_k}{m}}{\binom{n-1}{m}} - \frac{\binom{n_j}{m}}{\binom{n-1}{m}} + \frac{2n_jn_k}{(n_j+n_k)_{2\downarrow}} \frac{\binom{n_j+n_k}{m}}{\binom{n-1}{m}} \frac{1}{m+1} \right] \\
&= \frac{n_i}{n} \left\{ \frac{(n_j)_{2\downarrow}}{(n_j+n_k)_{2\downarrow}} \frac{n_j+n_k}{n_i} - \frac{n_j}{n_i+n_k} \right. \\
&\qquad \left. + 2\frac{n_jn_k}{(n_j+n_k)_{2\downarrow}} \left[ \frac{n}{n_j+n_k+1} (H_n - H_{n_i-1}) - 1 \right] \right\} \\
&= \frac{(n_j)_{2\downarrow}}{n(n_j+n_k-1)} - \frac{n_in_j}{n(n_i+n_k)} - 2\frac{n_in_jn_k}{n(n_j+n_k)_{2\downarrow}} \\
&\quad + 2\frac{n_in_jn_k}{(n_j+n_k+1)_{3\downarrow}} (H_n - H_{n_i-1}). \qquad (4.20)
\end{aligned}$$

Following a similar line of computation as above, we can find a closed-form expression for $\beta(\boldsymbol{n}, i, j, k)$ as follows:

$$\begin{aligned}
\beta(\boldsymbol{n}, i, j, k) &= \sum_{m=3}^{n} \frac{1}{m^2(m-1)} \sum_{\substack{\mathbf{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_i = 1, |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} m_j m_k \\
&= \sum_{m=3}^{n} \frac{1}{m^2(m-1)} \frac{\binom{n-n_i}{m-1}}{\binom{n}{m}} n_i \sum_{\substack{\mathbf{0} \prec \boldsymbol{m}' \preceq \boldsymbol{n} - n_i \boldsymbol{e}_i: \\ |\boldsymbol{m}'| = m-1}} \frac{\binom{\boldsymbol{n} - n_i \boldsymbol{e}_i}{\boldsymbol{m}'}}{\binom{n-n_i}{m-1}} m'_j m'_k \\
&= \sum_{m=3}^{n} \frac{1}{m^2(m-1)} \frac{\binom{n_j+n_k}{m-1}}{\binom{n}{m}} n_i \frac{n_j n_k}{(nj+n_k)_{2\downarrow}} (m-1)_{2\downarrow} \\
&= \sum_{m=1}^{n} \frac{n_i}{n} \frac{n_j n_k}{(nj+n_k)_{2\downarrow}} \frac{\binom{n_j+n_k}{m}}{\binom{n}{m}} \left(1 - \frac{2}{m+1}\right) \\
&= \frac{n_i}{n} \frac{n_j n_k}{(nj+n_k)_{2\downarrow}} \left\{ \frac{n_j + n_k}{n_i} - 2 \left[ \frac{n}{n_j + n_k + 1} (H_n - H_{n_i - 1}) - 1 \right] \right\} \\
&= \frac{n_j n_k}{n(n_j + n_k - 1)} + 2 \frac{n_i n_j n_k}{n(n_j + n_k)_{2\downarrow}} - 2 \frac{n_i n_j n_k}{(n_j + n_k + 1)_{3\downarrow}} (H_n - H_{n_i - 1}), \qquad (4.21)
\end{aligned}$$

where the second equality above is the same change of variables from $\boldsymbol{m}$ to $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{e}_i$ as in the $\alpha(\boldsymbol{n}, i, j, k)$ term. The third equality follows from identity (A.5) in Fact 5, and the second to last equality follows from Facts 1 and 3. Substituting (4.20) and (4.21) into (4.16), and using (4.1) gives

$$\begin{aligned}
Q(\boldsymbol{n}) &= \Lambda(\boldsymbol{n}) \sum_{i,j,k \text{ distinct}} \left[ \pi_j P_{ji} P_{jk} \alpha(\boldsymbol{n}, i, j, k) + \pi_k P_{kj} P_{ji} \beta(\boldsymbol{n}, i, j, k) \right] \\
&= \Lambda(\boldsymbol{n}) \sum_{i,j,k \text{ distinct}} \left\{ \pi_j P_{ji} P_{jk} \left[ \frac{(n_j)_{2\downarrow}}{n(n_j + n_k - 1)} - \frac{n_i n_j}{n(n_i + n_k)} - 2 \frac{n_i n_j n_k}{n(n_j + n_k)_{2\downarrow}} \right. \right. \\
&\qquad \left. + 2 \frac{n_i n_j n_k}{(n_j + n_k + 1)_{3\downarrow}} (H_n - H_{n_i - 1}) \right] \\
&\qquad + \pi_k P_{kj} P_{ji} \left[ \frac{n_j n_k}{n(n_j + n_k - 1)} + 2 \frac{n_i n_j n_k}{n(n_j + n_k)_{2\downarrow}} \right. \\
&\qquad \left. \left. - 2 \frac{n_i n_j n_k}{(n_j + n_k + 1)_{3\downarrow}} (H_n - H_{n_i - 1}) \right] \right\}. \qquad (4.22)
\end{aligned}$$

Note that if $\boldsymbol{P}$ is reversible when restricted to the observed alleles $\mathcal{O}_{\boldsymbol{n}}$, then (4.22) simplifies to the expression given in Corollary 5. □

## 4.6 Alternate proof of Theorem 4 via coupling

In this section, we give another proof of Theorem 4 by using a coupling argument with our urn construction described in Section 4.1. Suppose $\mathcal{O}_{\boldsymbol{n}} = \{a, b, c\}$ and $\boldsymbol{P}$ is not necessarily reversible restricted to $\mathcal{O}_{\boldsymbol{n}}$. As before, we consider killings in our urn process to take two time steps: one step in which we kill a color, and a second step in which we select a "parent" ball, create a copy of it, and add the copy to the urn. It turns out that our urn process would be much simpler to work with if, instead of duplicating the parent ball, we deleted it.

Let $\mathcal{U}$ be our original urn process. Define $\mathcal{U}''$ to be the urn process where we delete the parent ball, instead of duplicating it (so after each killing, we have two fewer balls than we would have under $\mathcal{U}$). At each time step of $\mathcal{U}''$, we remove a ball from the urn uniformly at random, so in $\mathcal{U}''$ we are simply sampling from the urn without replacement.

Define $\mathcal{U}'$ to be an "intermediate" process between $\mathcal{U}$ and $\mathcal{U}''$: on the first killing, we duplicate the parent ball as in $\mathcal{U}$, but on the second killing, we delete the parent ball as in $\mathcal{U}''$. Since $|\mathcal{O}_{\boldsymbol{n}}| = 3$, there will only be two killing events before the end of the urn process.

The first thing to note is that the probability of drawing a tree $T$ is the same under $\mathcal{U}$ and $\mathcal{U}'$. This is because when the second killing happens, $T$ will have already been determined, and so it does not matter whether we duplicate or delete the parent ball. Henceforth, we will work with the process $\mathcal{U}'$ instead of the process $\mathcal{U}$ to compute $\mathbb{P}_{\boldsymbol{n}}(T)$. The next thing to note is that the probability of drawing $T$ under $\mathcal{U}''$ is straightforward to compute. So we will compute these probabilities under $\mathcal{U}''$, and then use a coupling between $\mathcal{U}'$ and $\mathcal{U}''$ to compute $\mathbb{P}_{\boldsymbol{n}}(T)$.

Denote $\mathcal{U}''(\boldsymbol{n})$ as the process $\mathcal{U}''$ with starting configuration $\boldsymbol{n}$. We will use the following two lemmas for $\mathcal{U}''(\boldsymbol{n})$:

**Lemma 9.** *In $\mathcal{U}''(\boldsymbol{n})$, consider any particular ball of color $k$, and condition on this ball being the last one deleted. Under this conditioning, the previous $n-1$ choices follow the law $\mathcal{U}''(\boldsymbol{n} - \boldsymbol{e}_k)$.*

*Proof.* Label the balls in the urn, and let $\sigma_i$ be the $i$th ball we delete. Let $A_j$ be the event that we delete ball $j$ last. Then, for permutation $\sigma$ with $\sigma_n = j$,

$$\mathbb{P}((\sigma_1, \cdots, \sigma_{n-1}) \mid A_j) = \frac{\mathbb{P}(\sigma)}{\mathbb{P}(\sigma_n = j)} = \frac{1/n!}{1/n} = \frac{1}{(n-1)!}. \qquad \square$$

**Lemma 10.** *In $\mathcal{U}''(\boldsymbol{n})$, consider any particular ball in the urn, and define $Y_t$ to be $0$ if it is deleted before time $t+1$, and $1/(n-t)$ if it is not (i.e., $Y_t$ is the total proportion of mass in the urn that this ball contributes after the $t$-th draw). Let $\mathcal{F}_t$ be the $\sigma$-algebra generated by all sequences of choices up to time $t$. Then $\{(Y_t, \mathcal{F}_t), t \geq 0\}$ is a martingale.*

*Proof.* If $Y_{t-1} = 0$, then $Y_t = 0$ as well. Otherwise,

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = \frac{n-t}{n-t+1} \cdot \frac{1}{n-t} = Y_{t-1}. \qquad \square$$

We now compute $\mathbb{P}(T;\mathcal{U}''(\boldsymbol{n}))$, the probability of drawing $T$ under $\mathcal{U}''(\boldsymbol{n})$. Let $T^{\rho}_{ijk}$ be the tree whose interior vertex is $j$, with root at $\rho$. Let $D_i$ be the event where the first color killed is $i$.

First, consider $\mathbb{P}(T^k_{ijk};\mathcal{U}''(\boldsymbol{n})) = \mathbb{P}(T^k_{ijk} \cap D_i;\mathcal{U}''(\boldsymbol{n}))$. Since the sequence of draws from $\mathcal{U}''(\boldsymbol{n})$ is exchangeable, the probability that any particular ball of color $k$ is deleted last is $1/n$. Conditional on this, the probability that any particular ball of color $j$ is the parent of color $i$ is $1/(n-1)$. To see this, note that from Lemma 9, the first $n-1$ steps of $U''(\boldsymbol{n})$ conditioned on a particular ball of color $k$ being deleted last follows the law of $\mathcal{U}''(\boldsymbol{n}-\boldsymbol{e}_k)$. In $\mathcal{U}''(\boldsymbol{n}-\boldsymbol{e}_k)$, let $Y_t$ be the proportion of mass of the ball of color $j$, and let $\tau$ be the time that $i$ is killed. From Lemma 10 and the tower property, the probability that the ball of color $j$ is the parent of $i$ is $\mathbb{E}[Y_\tau] = Y_0 = 1/(n-1)$ Hence

$$\mathbb{P}(T^k_{ijk};\mathcal{U}''(\boldsymbol{n})) = \mathbb{P}(T^k_{ijk} \cap D_i;\mathcal{U}''(\boldsymbol{n})) = \frac{n_k}{n} \cdot \frac{n_j}{n-1}. \tag{4.23}$$

Next, we find $\mathbb{P}(T^k_{ikj} \cap D_i;\mathcal{U}''(\boldsymbol{n}))$. Note that the probability of first killing color $i$, and then killing color $j$ is

$$\mathbb{P}((T^k_{ikj} \cup T^k_{ijk}) \cap D_i;\mathcal{U}''(\boldsymbol{n})) = \frac{n_k}{n} \cdot \frac{n_j}{n_i+n_j}. \tag{4.24}$$

To see this, note that the probability of any particular ball of color $k$ being deleted last is $1/n$. Conditional on this, the probability that any particular ball of color $i$ or $j$ is the last ball of these two colors to be deleted is $1/(n_i+n_j)$, which can be seen as follows: by construction of $\mathcal{U}''(\boldsymbol{n})$, the subsequence of draws that do not involve a ball of color $k$ follows the law of $\mathcal{U}''(\boldsymbol{n}-n_k\boldsymbol{e}_k)$, and the probability of deleting any particular ball of color $i$ or $j$ last in $\mathcal{U}''(\boldsymbol{n}-n_k\boldsymbol{e}_k)$ is $1/(n_i+n_j)$. Hence (4.24) holds. Finally, subtracting (4.23) from (4.24) gives us

$$\mathbb{P}(T^k_{ikj} \cap D_i;\mathcal{U}''(\boldsymbol{n})) = \frac{n_k n_j (n-1-n_i-n_j)}{n(n_i+n_j)(n-1)} = \frac{n_k(n_k-1)n_j}{n(n-1)(n_i+n_j)}. \tag{4.25}$$

We now compute $\mathbb{P}_{\boldsymbol{n}}(T)$, the probability of drawing $T$ under the process $\mathcal{U}'$. Recall that $\mathbb{P}_{\boldsymbol{n}}(T)$ is the same under $\mathcal{U}$ and $\mathcal{U}'$. The only difference between $\mathcal{U}'$ and $\mathcal{U}''$ is that $\mathcal{U}'$ has two additional balls after the first killing, because we duplicate the parent ball in $\mathcal{U}'$ but delete it in $\mathcal{U}''$. So we set up a coupling between $\mathcal{U}'$ and $\mathcal{U}''$ by running $\mathcal{U}'$ normally. To get the coupled process $\mathcal{U}''$, ignore the two extra balls we have after the first killing. When we select one of these extra balls, we pretend that this did not happen and do not count it as a move in $\mathcal{U}''$. To see that the coupled process $\mathcal{U}''$ has the correct distribution, note that each ball in the coupled process $\mathcal{U}''$ has an equal chance of being the next removed. Hence the transition probabilities are correct and $\mathcal{U}''$ has the correct distribution.

We now compute $\mathbb{P}_{\boldsymbol{n}}(T^b_{abc} \cap D_a)$ and $\mathbb{P}_{\boldsymbol{n}}(T^c_{abc} \cap D_a)$. Let $E_i = T^i_{abc} \cap D_a$, and let $E_* = E_b \cup E_c$. In words, $E_*$ is the event where we kill $a$ first and select $b$ as its parent, while $E_i$ is the event that we kill $a$ first and kill $i$ last. Note that $E_*$ occurs in $\mathcal{U}'$ if and only if it occurs in $\mathcal{U}''$, since under the coupling, $\mathcal{U}'$ and $\mathcal{U}''$ are identical up to the first killing.

If event $A$ happens under $\mathcal{U}'$ and event $B$ happens under $\mathcal{U}''$ in the coupling, we denote this joint event as $A \times B$. Let $\Omega_{\mathcal{U}'}$ be the sample space of $\mathcal{U}'$ (likewise for $\Omega_{\mathcal{U}''}$ and $\mathcal{U}''$). The observation above yields the following equalities for all $A \subseteq E_*$:

$$\mathbb{P}(A \times \Omega_{\mathcal{U}''}) = \mathbb{P}(A \times E_*), \tag{4.26}$$

$$\mathbb{P}(\Omega_{\mathcal{U}'} \times A) = \mathbb{P}(E_* \times A). \tag{4.27}$$

Next, note that if $E_c$ occurs in $\mathcal{U}'$, then this implies that it occurs in $\mathcal{U}''$. For if we kill color $c$ last in $\mathcal{U}'$, then color $c$ will still be killed last in $\mathcal{U}''$ if we ignore the two extra balls of color $b$. Similarly, if $E_b$ occurs in $\mathcal{U}''$, then $E_b$ must also occur in $\mathcal{U}'$. For if we kill color $b$ last in $\mathcal{U}''$, then color $b$ will still be killed last in $\mathcal{U}'$ if we insert two extra balls of color $b$ after the first killing. Summarizing these observations, we have

$$\mathbb{P}(E_c \times \Omega_{\mathcal{U}''}) = \mathbb{P}(E_c \times E_c) \tag{4.28}$$

$$\mathbb{P}(\Omega_{\mathcal{U}'} \times E_b) = \mathbb{P}(E_b \times E_b). \tag{4.29}$$

Using (4.27) and (4.28), we get

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{n}}(T^c_{abc} \cap D_a) &= \mathbb{P}(E_c \times \Omega_{\mathcal{U}''}) \\
&= \mathbb{P}(E_c \times E_c) \\
&= \mathbb{P}(E_* \times E_c) - \mathbb{P}(E_b \times E_c) \\
&= \mathbb{P}(\Omega_{\mathcal{U}'} \times E_c) - \mathbb{P}(E_b \times E_c) \\
\mathbb{P}_{\boldsymbol{n}}(T^c_{abc} \cap D_a) &= \mathbb{P}(T^c_{abc} \cap D_a; \mathcal{U}''(\boldsymbol{n})) - \mathbb{P}(E_b \times E_c).
\end{aligned} \tag{4.30}$$

Similarly, using (4.26) and (4.29) yields

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{n}}(T^b_{abc} \cap D_a) &= \mathbb{P}(E_b \times \Omega_{\mathcal{U}''}) \\
&= \mathbb{P}(E_b \times E_*) \\
&= \mathbb{P}(E_b \times E_b) + \mathbb{P}(E_b \times E_c) \\
&= \mathbb{P}(\Omega_{\mathcal{U}'} \times E_b) + \mathbb{P}(E_b \times E_c) \\
\mathbb{P}_{\boldsymbol{n}}(T^b_{abc} \cap D_a) &= \mathbb{P}(T^b_{abc} \cap D_a; \mathcal{U}''(\boldsymbol{n})) + \mathbb{P}(E_b \times E_c).
\end{aligned} \tag{4.31}$$

Now, we have already computed $\mathbb{P}(T^b_{abc} \cap D_a; \mathcal{U}''(\boldsymbol{n}))$ and $\mathbb{P}(T^c_{abc} \cap D_a; \mathcal{U}''(\boldsymbol{n}))$ in (4.23) and (4.25). So the only thing remaining is to compute $\mathbb{P}(E_b \times E_c)$.

If $E_b \times E_c$ occurs, then color $b$ is killed last under $\mathcal{U}'$, but color $c$ is killed last under $\mathcal{U}''$. Since $\mathcal{U}'$ is identical to $\mathcal{U}''$ except for two extra balls of color $b$, this means that one of these extra balls in $\mathcal{U}'$ is the last to be removed, while among the balls in $\mathcal{U}''$, color $c$ is killed last. Applying Lemmas 9 and 10 again, the conditional probability of this happening is

$$\frac{2}{m} \cdot \frac{m_c}{m-2} \tag{4.32}$$

where we condition on having sample $\mathbf{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}$ right before drawing the edge $(b \to a)$. To see this, note that we must select one of the extra balls of color $b$ last and this happens with

probability $2/m$. Conditional on this, we use the martingale property to find the probability that a ball of color $c$ is the parent of the non-extra balls of color $b$, which happens with probability $m_c/(m-2)$. Finally, we use (4.13) to compute $\mathbb{P}(E_b \times E_c)$. While (4.13) was shown for $\mathcal{U}$, repeating the argument but substituting $\mathcal{U}'$ or $\mathcal{U}''$ shows that it holds for these urn processes as well. This is because the proof of (4.13) only depends on the sequence of draws up to the first killing, and up to this point $\mathcal{U}$, $\mathcal{U}'$, and $\mathcal{U}''$ are identically distributed. Hence, by (4.13) and (4.32), we have

$$\begin{aligned}
\mathbb{P}(E_b \times E_c) &= \sum_{\substack{\mathbf{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ m_a = 1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_b}{m(m-1)} \frac{2m_c}{m(m-2)} \\
&= 2 \sum_{m=3}^{n-n_a+1} \frac{n_a}{n} \frac{\binom{n_b+n_c}{m-1}}{\binom{n-1}{m-1}} \frac{n_b n_c}{(n_b+n_c)_{2\downarrow}} \frac{1}{m} \\
&= \frac{2n_a n_b n_c}{n(n_b+n_c)_{2\downarrow}} \left( \frac{n}{n_b+n_c+1}(H_n - H_{n_a-1}) - 1 - \frac{1}{2}\frac{n_b+n_c}{n-1} \right) \\
\mathbb{P}(E_b \times E_c) &= \frac{2n_a n_b n_c}{(n_b+n_c+1)_{3\downarrow}}(H_n - H_{n_a-1}) - \frac{2n_a n_b n_c}{n(n_b+n_c)_{2\downarrow}} - \frac{n_a n_b n_c}{n(n-1)(n_b+n_c-1)},
\end{aligned}$$

where the second equality is due to Fact 4 and the third equality is due to Fact 2. Combining this with (4.23), (4.25), (4.30) and (4.31) yields

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{n}}(T^b_{abc} \cap D_a) &= \frac{(n_b)_{2\downarrow} n_c}{(n)_{2\downarrow}(n_a+n_c)} + \mathbb{P}(E_b \times E_c) \\
\mathbb{P}_{\boldsymbol{n}}(T^c_{abc}) &= \frac{n_c n_b}{(n)_{2\downarrow}} - \mathbb{P}(E_b \times E_c),
\end{aligned}$$

which simplify to $\alpha(\boldsymbol{n}, a, b, c)$ and $\beta(\boldsymbol{n}, a, b, c)$ respectively in Section 4.5. □

## 4.7 Proof of Theorem 6 ($|\mathcal{O}_{\boldsymbol{n}}| = 4$)

Using Corollary 5, we first note the following alternate expression for $R(\boldsymbol{n})$ when $|\mathcal{O}_{\boldsymbol{n}}| = 3$ and $\boldsymbol{P}$ is reversible restricted to the observed alleles:

$$R(\boldsymbol{n}) = \sum_{i,j,k \text{ distinct}} \frac{n_i}{n} \pi_i \frac{P_{ij} P_{ik}}{2}. \tag{4.33}$$

Suppose $|\mathcal{O}_{\boldsymbol{n}}| = 4$ and assume that $\boldsymbol{P}$ is reversible restricted to the observed alleles $\mathcal{O}_{\boldsymbol{n}}$. Then using Proposition 8, we obtain

$$\begin{aligned}
R(\boldsymbol{n}) &= \sum_{l,h\neq l} P_{hl} \sum_{\substack{\mathbf{0}\prec\boldsymbol{m}\preceq\boldsymbol{n}:\\ m_l=1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_h R(\boldsymbol{m}-\boldsymbol{e}_l+\boldsymbol{e}_h)}{m(m-1)} \\
&= \sum_{l,h\neq l} P_{hl} \sum_{\substack{\mathbf{0}\prec\boldsymbol{m}\preceq\boldsymbol{n}:\\ m_l=1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_h}{m(m-1)} \sum_{\substack{i,j,k \text{ distinct},\\ i,j,k\neq l}} \frac{m_i+\delta_{i,h}}{m}\pi_i \frac{P_{ij}P_{ik}}{2} \\
&= \sum_{i,j,k,l \text{ distinct}} \frac{1}{2}\pi_i P_{ij}P_{ik}P_{il} \sum_{\substack{\mathbf{0}\prec\boldsymbol{m}\preceq\boldsymbol{n}\\ m_l=1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_i(m_i+1)}{m^2(m-1)} \\
&\quad + \sum_{i,j,k,l \text{ distinct}} \pi_i P_{ij}P_{ik}P_{jl} \sum_{\substack{\mathbf{0}\prec\boldsymbol{m}\preceq\boldsymbol{n}\\ m_l=1}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \frac{m_i m_j}{m^2(m-1)}, \qquad (4.34)
\end{aligned}$$

where the second equality follows from using (4.33) since $\boldsymbol{P}$ is reversible when restricted to the alleles $\{i,j,k\} \subset \mathcal{O}_{\boldsymbol{n}}$. Similar to the proof in Section 4.5, if we define quantities $\zeta(\boldsymbol{n},i,j,k,l)$ and $\delta(\boldsymbol{n},i,j,k,l)$ as

$$\zeta(\boldsymbol{n},i,j,k,l) = \sum_{m=4}^{n} \frac{1}{m^2(m-1)} \sum_{\substack{\mathbf{0}\prec\boldsymbol{m}\preceq\boldsymbol{n}:\\ m_l=1,|\boldsymbol{m}|=m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} m_i(m_i+1),$$

and

$$\delta(\boldsymbol{n},i,j,k,l) = \sum_{m=4}^{n} \frac{1}{m^2(m-1)} \sum_{\substack{\mathbf{0}\prec\boldsymbol{m}\preceq\boldsymbol{n}:\\ m_l=1,|\boldsymbol{m}|=m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} m_i m_j,$$

then, using (4.1) and (4.34), we obtain the following expression for $Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n})R(\boldsymbol{n})$:

$$Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n}) \sum_{i,j,k,l \text{ distinct}} \left[ \pi_i P_{ij}P_{ik}P_{il} \frac{\zeta(\boldsymbol{n},i,j,k,l)}{2} + \pi_i P_{ij}P_{ik}P_{jl}\delta(\boldsymbol{n},i,j,k,l) \right]. \qquad (4.35)$$

By a very similar calculation to that in Section 4.5, using Facts 1 and 3, and identities (A.4) and (A.5) in Fact 5 of the Appendix, we obtain the following closed-form expressions for

$\zeta(\boldsymbol{n}, i, j, k, l)$ and $\delta(\boldsymbol{n}, i, j, k, l)$:

$$\begin{aligned}
&\zeta(\boldsymbol{n}, i, j, k, l) \\
&\quad = \frac{n_l}{n}\left\{\frac{n_i+n_j+n_k}{n_l}\frac{(n_i)_{2\downarrow}}{(n_i+n_j+n_k)_{2\downarrow}} + \frac{n_i}{n_j+n_k+n_l}\right. \\
&\qquad + \frac{2n_i(n_j+n_k)}{(n_i+n_j+n_k)_{2\downarrow}}\left(\frac{n}{n_i+n_j+n_k+1}(H_n - H_{n_l-1}) - 1\right) \\
&\qquad - \left[\frac{n_i+n_j}{n_k+n_l}\frac{(n_i)_{2\downarrow}}{(n_i+n_j)_{2\downarrow}} + \frac{2n_in_j}{(n_i+n_j)_{2\downarrow}}\left(\frac{n}{n_i+n_j+1}(H_n - H_{n_k+n_l-1}) - 1\right)\right] \\
&\qquad \left. - \left[\frac{n_i+n_k}{n_j+n_l}\frac{(n_i)_{2\downarrow}}{(n_i+n_k)_{2\downarrow}} + \frac{2n_in_k}{(n_i+n_k)_{2\downarrow}}\left(\frac{n}{n_i+n_k+1}(H_n - H_{n_j+n_l-1}) - 1\right)\right]\right\}.
\end{aligned}$$

and

$$\begin{aligned}
&\delta(\boldsymbol{n}, i, j, k, l) \\
&\quad = \frac{n_l}{n}\left\{\frac{n_i+n_j+n_k}{n_l}\frac{n_in_j}{(n_i+n_j+n_k)_{2\downarrow}}\right. \\
&\qquad - \frac{2n_in_j}{(n_i+n_j+n_k)_{2\downarrow}}\left(\frac{n}{n_i+n_j+n_k+1}(H_n - H_{n_l-1}) - 1\right) \\
&\qquad \left. - \left[\frac{n_i+n_j}{n_k+n_l}\frac{n_in_j}{(n_i+n_j)_{2\downarrow}} - \frac{2n_in_j}{(n_i+n_j)_{2\downarrow}}\left(\frac{n}{n_i+n_j+1}(H_n - H_{n_k+n_l-1}) - 1\right)\right]\right\}.
\end{aligned}$$

Simplifying the expression for $\delta(\boldsymbol{n}, i, j, k, l)$, we get the expression stated in Theorem 6. Observing that $\zeta(\boldsymbol{n}, i, j, k, l)$ is symmetric in $j$ and $k$, we see that for all $i, j, k$, and $l$ distinct in $\mathcal{O}_{\boldsymbol{n}}$,

$$\frac{\zeta(\boldsymbol{n}, i, j, k, l) + \zeta(\boldsymbol{n}, i, k, j, l)}{2} = \gamma(\boldsymbol{n}, i, j, k, l) + \gamma(\boldsymbol{n}, i, k, j, l),$$

where $\gamma(\boldsymbol{n}, i, j, k, l)$ is given by:

$$\begin{aligned}
\gamma(\boldsymbol{n}, i, j, k, l) = \frac{n_i}{n}\Bigg\{&\left[\frac{n_i-1}{2(n_i+n_j+n_k-1)} - \frac{2n_jn_l}{(n_i+n_j+n_k)_{2\downarrow}}\right] + \frac{n_l}{2(n_j+n_k+n_l)} \\
&- \left[\frac{n_l(n_i-1)}{(n_k+n_l)(n_i+n_j-1)} - \frac{2n_jn_l}{(n_i+n_j)_{2\downarrow}}\right]\Bigg\} \\
&+ \frac{2n_in_jn_l}{(n_i+n_j+n_k+1)_{3\downarrow}}(H_n - H_{n_l-1}) - \frac{2n_in_jn_l}{(n_i+n_j+1)_{3\downarrow}}(H_n - H_{n_k+n_l-1}).
\end{aligned}$$

Using the fact that $\pi_iP_{ij}P_{ik}P_{il}$ is also symmetric in $j$ and $k$, we can then rewrite (4.35) as

$$Q(\boldsymbol{n}) = \Lambda(\boldsymbol{n}) \sum_{i,j,k,l \text{ distinct}} \Big(\pi_iP_{ij}P_{ik}P_{il}\gamma(\boldsymbol{n}, i, j, k, l) + \pi_iP_{ij}P_{ik}P_{jl}\delta(\boldsymbol{n}, i, j, k, l)\Big). \qquad \square$$

# Chapter 5

# Numerical evaluation of accuracy

In this chapter, we investigate the accuracy of approximating the sampling probability $q(\boldsymbol{n})$ by using only the leading order term $\theta^{|\mathcal{O}_{\boldsymbol{n}}|-1}Q(\boldsymbol{n})$. In order to compare the accuracy of our approximate formulas against the exact sampling probability, we solve the recursion (2.6) numerically to obtain the true sampling probability $q(\boldsymbol{n})$.

For a given sample $\boldsymbol{n}$, define the approximate sampling probability, $q_{\text{approx}}(\boldsymbol{n})$, by

$$q_{\text{approx}}(\boldsymbol{n}) = \theta^{|\mathcal{O}_{\boldsymbol{n}}|-1}Q(\boldsymbol{n}).$$

We can then define the relative error, $\text{Err}(\boldsymbol{n})$, of the approximation $q_{\text{approx}}(\boldsymbol{n})$ from the true sampling probability $q(\boldsymbol{n})$ as

$$\text{Err}(\boldsymbol{n}) = \frac{|q(\boldsymbol{n}) - q_{\text{approx}}(\boldsymbol{n})|}{q(\boldsymbol{n})}.$$

For a given sample size $n$, another natural measure of the approximation quality is the expected relative error under the distribution arising from the coalescent on samples of size $n$. Since $q(\boldsymbol{n})$ is the probability of a particular ordered sample consistent with $\boldsymbol{n}$, the probability $p(\boldsymbol{n})$ of the unordered sample $\boldsymbol{n}$, when sampling order is ignored, is given by

$$p(\boldsymbol{n}) = \binom{n}{n_1, \ldots, n_K} q(\boldsymbol{n}).$$

We can then define the expected relative error for a sample size $n$ by $\text{AvgErr}(n)$, given by

$$\text{AvgErr}(n) = \sum_{\boldsymbol{n}:|\boldsymbol{n}|=n} p(\boldsymbol{n})\,\text{Err}(\boldsymbol{n}) = \sum_{\boldsymbol{n}:|\boldsymbol{n}|=n} \binom{n}{n_1, \ldots, n_K} |q(\boldsymbol{n}) - q_{\text{approx}}(\boldsymbol{n})|.$$

We also define the worst-case relative error, $\text{WorstErr}(n)$, for a given sample size $n$ as the worse relative error among all samples of size $n$. Specifically,

$$\text{WorstErr}(n) = \max_{\boldsymbol{n}:|\boldsymbol{n}|=n} \text{Err}(\boldsymbol{n}) = \max_{\boldsymbol{n}:|\boldsymbol{n}|=n} \frac{|q(\boldsymbol{n}) - q_{\text{approx}}(\boldsymbol{n})|}{q(\boldsymbol{n})}.$$

To study the accuracy of approximating $q(\boldsymbol{n})$ by $q_{\text{approx}}(\boldsymbol{n})$, we examine the behavior of AvgErr$(n)$ and WorstErr$(n)$ for a transition matrix estimated from real biological data. Specifically, we use the reversible phylogenetic mutation rate matrix estimated in [48, Table 1, matrix (1)] for the $\psi\eta$-globin pseudogenes of six primate species. Since their estimated matrix is a matrix of nucleotide substitution rates used for phylogenetic analysis, we rescale it by the minimum amount that can make it a valid Markov transition matrix. This rescaled matrix, denoted by $\widehat{\boldsymbol{P}}$, is given below to three digits of precision, and is used in our numerical experiments with different values of the mutation parameter $\theta$:

$$\widehat{\boldsymbol{P}} = \begin{pmatrix} 0.433 & 0.398 & 0.074 & 0.095 \\ 0.665 & 0.000 & 0.164 & 0.171 \\ 0.074 & 0.098 & 0.394 & 0.434 \\ 0.147 & 0.159 & 0.674 & 0.020 \end{pmatrix}, \tag{5.1}$$

in the $(T, C, A, G)$ basis. The stationary distribution corresponding to this transition matrix is $\widehat{\boldsymbol{\pi}} = (0.308, 0.185, 0.308, 0.199)$ to three digits of precision.

For many neutral regions of the human genome, typical mutation rates per base are in the range $10^{-3} \leq \theta \leq 10^{-2}$ [35], and we consider $\theta \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$ in our study. For the transition matrix in (5.1), the expected relative error AvgErr$(n)$ and the worst-case relative error WorstErr$(n)$ are plotted in Figures 5.1 and 5.2, respectively, as functions of the sample size $n$. As can be seen from the plots, both the expected relative error and the worst-case relative error grow very slowly with the sample size $n$. Further, the ratio of WorstErr$(n)$ to AvgErr$(n)$ is a small number between 1.3 and 2.1 for all $n \leq 360$, and is decreasing in $n$. Hence, it appears that the approximation quality of $q_{\text{approx}}(\boldsymbol{n})$ is uniformly good over all samples $\boldsymbol{n}$ for any given size $n$.

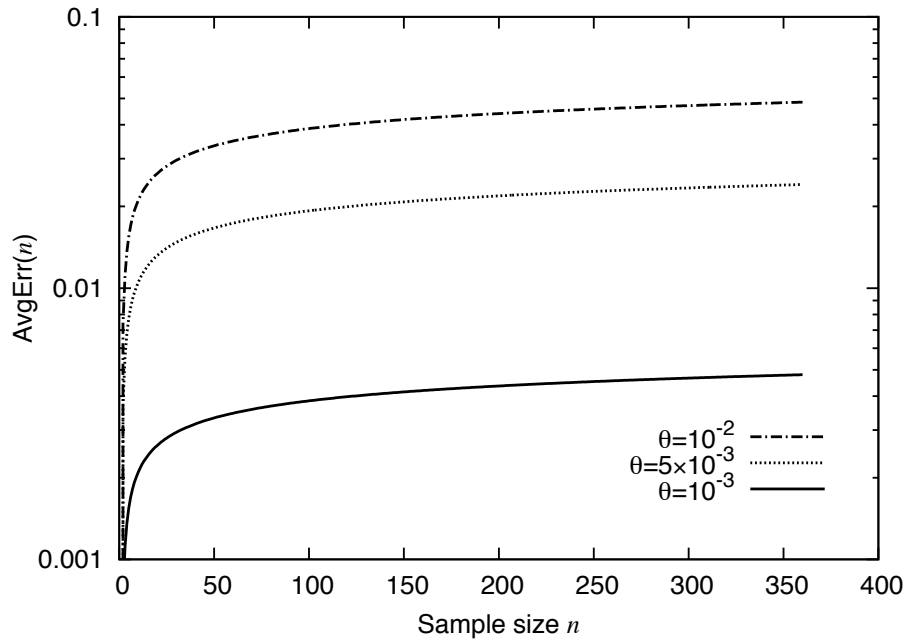


Figure 5.1: The expected relative error, AvgErr($n$), as a function of the sample size $n$, for the transition matrix $\widehat{\boldsymbol{P}}$ in (5.1) and mutation rate $\theta \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$.

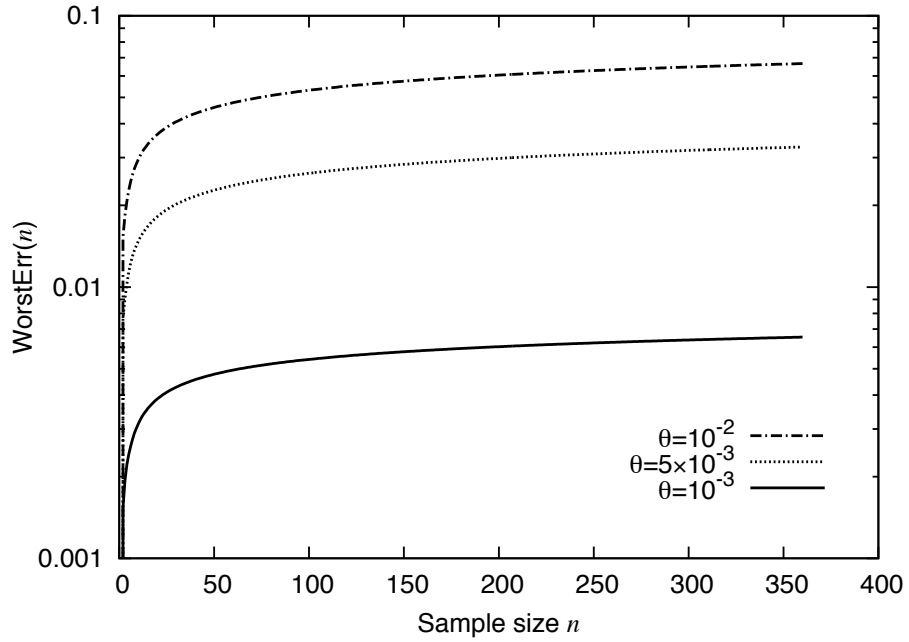


Figure 5.2: The worst-case relative error, WorstErr($n$), as a function of the sample size $n$, for the transition matrix $\widehat{\boldsymbol{P}}$ in (5.1) and mutation rate $\theta \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$.

# Chapter 6

# Discussion

In this thesis, we have developed explicit expressions for approximating the probability of sampling a given configuration of alleles under the coalescent with general finite-alleles models of mutation, when up to four distinct alleles are observed in the sample. These formulas are accurate up to leading order in the mutation rate $\theta$, and they can be evaluated in constant time and space using some precomputation. Furthermore, numerical experiments suggest that these formulas are highly accurate for the mutation rates that are of relevance to SNPs in humans.

Our results and proofs crucially depend on the population being at stationarity, i.e. the population size is constant over time, and the allelic type of the most recent common ancestor of the sample is drawn from the stationary distribution of the mutation matrix. It is also interesting to consider the problem of determining exact or approximate sampling probabilities for time-varying population size models. For example, various studies have found evidence for severe bottlenecks in the human effective population size during the dispersal out of Africa around 60,000 years ago (see [28] for a survey). Recent large-sample sequencing studies also predict an exponential growth in the human effective population size in the last several hundred generations [4, 36, 44]. One method for inferring such historical population size changes is based on computing the probability of a sample configuration at a polymorphic locus, conditional on the demographic model and the observed locus being polymorphic [30, 4]. For general time-varying population sizes, formulas for the first-order approximation in the mutation rate $\theta$ of this conditional probability are known when two alleles are observed at the polymorphic locus [39, 40]. However, no such leading-order approximations are known when three or more distinct alleles are observed in the sample and when the population sizes are general functions varying over time, even in the case of the infinite-alleles or finite-alleles PIM models. Even if it is difficult to supply succinct closed-form expressions, it would be interesting to develop efficient algorithms for computing the sampling probability for large samples under variable population size models. Such algorithms would have great applicability to demographic inference, where large sample sizes need to be analyzed in order to infer rapid population growth in recent history [23].

Another interesting perspective on the problem of computing sampling probabilities

would be to consider the family of Wright-Fisher diffusion processes that are dual to the coalescent family of models. The sampling probability under the coalescent can be expressed in terms of the stationary measure of the Wright-Fisher diffusion. More precisely, if $\psi(\boldsymbol{x}) = \psi(x_1, \ldots, x_K)$ is the stationary measure of a Wright-Fisher diffusion for a $K$-allelic mutation model with mutation matrix $\boldsymbol{P}$, the ordered sampling probability $q(\boldsymbol{n} \mid \theta, \boldsymbol{P})$ is given by

$$q(\boldsymbol{n} \mid \theta, \boldsymbol{P}) = \int_{\Delta^{k-1}} \left( \prod_{i=1}^{K} x_i^{n_i} \right) \psi(\boldsymbol{x}) d\boldsymbol{x}, \tag{6.1}$$

where $\Delta^{k-1} = \{(x_1, \ldots, x_K) : x_i \geq 0 \ \forall i, \ \sum_{i=1}^{K} x_i = 1\}$ is the $(K-1)$-dimensional simplex. For the Wright-Fisher diffusion with a finite-alleles PIM matrix, the stationary measure $\psi$ is given by the Dirichlet distribution with parameters $\theta\pi_1, \ldots, \theta\pi_K$. Using (6.1) with this stationary measure, one can recover Wright's sampling formula for PIM matrices given in (2.5). However, no analytic expression is known for the stationary measure $\psi$ if $\boldsymbol{P}$ is a general mutation transition matrix. Hence, it would be interesting to find an approximate stationary measure for the diffusion with general mutation matrices, such that using this approximation for $\psi$ in (6.1) would recover asymptotically accurate sampling formulas. Even if turns out to be difficult to evaluate the integral in (6.1) in closed-form, it might be possible to efficiently estimate the value of the integral using numerical methods.

A third direction of future research would be to consider an asymptotic expansion for the sampling probability for large mutation rates. This could be relevant for microsatellite loci where the population-scaled mutation rate $\theta \gg 1$. Considering such an expansion,

$$q(\boldsymbol{n} \mid \theta, \boldsymbol{P}) = q^{(0)}(\boldsymbol{n} \mid \boldsymbol{P}) + \frac{q^{(1)}(\boldsymbol{n} \mid \boldsymbol{P})}{\theta} + \frac{q^{(2)}(\boldsymbol{n} \mid \boldsymbol{P})}{\theta^2} + \cdots, \tag{6.2}$$

it is easy to see that the zeroth-order coefficient $q^{(0)}$ is given by

$$q^{(0)}(\boldsymbol{n} \mid \boldsymbol{P}) = \prod_{i=1}^{K} \pi_i^{n_i}, \tag{6.3}$$

where $\boldsymbol{\pi}$ is the stationary distribution of $\boldsymbol{P}$. It would be interesting to determine closed-form expressions or efficient algorithms for computing higher-order coefficients $q^{(i)}$ for $i \geq 1$.

It is worth mentioning that even though the formulas derived in this thesis directly apply to sample configurations at a single locus, they could also have applications to haplotype configurations at two or more loci. The reason is as follows: in multi-locus models with finite recombination rates, no closed-form sampling formula is known, even for the simplest case of two loci with either infinite-alleles or finite-alleles PIM models. However, a new framework based on asymptotic expansions has recently been developed [19, 20, 22, 3] to derive useful closed-form approximations when the recombination rate is moderate to large by performing an asymptotic expansion of the sampling probability in inverse powers of

the recombination rate. We note that our one-locus sampling formula for the $|\mathcal{O}_{\boldsymbol{n}}| = 4$ case provides an accurate approximation of the sampling probability for a completely linked (i.e., with zero recombination rate) pair of loci with two observed alleles at each locus (as is typical in SNP data). Hence, our formulas serve as a starting point for finding approximate two-locus sampling formulas when the recombination rate is small, complementary to earlier work [19, 20, 22, 3] for large recombination rates.

# Bibliography


[1] Arratia, A., Barbour, A. D. and Tavaré, S. (2003). *Logarithmic Combinatorial Structures: A Probabilistic Approach*. European Mathematical Society Publishing House, Switzerland.

[2] Bhaskar, A., Kamm, J. A. and Song, Y. S. (2012). Approximate sampling formulae for general finite-alleles models of mutation. *Advances in Applied Probability* **44,** 408–428.

[3] Bhaskar, A. and Song, Y. S. (2012). Closed-form asymptotic sampling distributions under the coalescent with recombination for an arbitrary number of loci. *Advances in Applied Probability* **44,** 391–407.

[4] Coventry, A., Bull-Otterson, L. M., Liu, X. et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications* **1,** 131.

[5] Durrett, R. (2008). *Probability models for DNA sequence evolution*. Springer.

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

[7] Fisher, R. (1930). *The genetical theory of natural selection*. Clarendon Press, Oxford.

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

[9] Griffiths, R. (1991). The two-locus ancestral graph. *Selected Proceedings of the Sheffield Symposium on Applied Probability. IMS Lecture Notes–Monograph Series* **18,** 100–117.

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

[11] Griffiths, R. and Lessard, S. (2005). Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theoretical Population Biology* **68,** 167–77.

[12] Griffiths, R. and Marjoram, P. (1996). Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology* **3,** 479–502.

[13] Griffiths, R. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **344,** 403–410.

[14] Griffiths, R. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46,** 131–159.

[15] Griffiths, R. and Tavaré, S. (1999). The ages of mutations in gene trees. *Annals of Applied Probability* 567–590.

[16] Herbots, H. (1997). The structured coalescent. *Progress in Population Genetics and Human Evolution* **(IMA Vol. Math. Appl. 87),** 231–255.

[17] Hoppe, F. (1984). Pólya-like urns and the Ewens' sampling formula. *Journal of Mathematical Biology* **20,** 91–94.

[18] Hudson, R. (2001). Two-locus sampling distributions and their application. *Genetics* **159,** 1805–1817.

[19] Jenkins, P. A. and Song, Y. S. (2009). Closed-form two-locus sampling distributions: Accuracy and universality. *Genetics* **183,** 1087–1103.

[20] Jenkins, P. A. and Song, Y. S. (2010). An asymptotic sampling formula for the coalescent with recombination. *The Annals of Applied Probability* **20,** 1005–1028.

[21] Jenkins, P. A. and Song, Y. S. (2011). The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theoretical Population Biology,* **80,** 158–173.

[22] Jenkins, P. A. and Song, Y. S. (2012). Padé approximants and exact two-locus sampling distributions. *The Annals of Applied Probability* **22,** 576–607.

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

[24] Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications* **13,** 235–248.

[25] Kingman, J. F. C. (1982). Exchangeability and the evolution of large populations. In *Exchangeability in probability and statistics*. ed. G. Koch and F. Spizzichino. North-Holland Publishing Company pp. 97–112.

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

[27] Krone, S. M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51,** 210–237.

[28] Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496.

[29] Lundstrom, R., Tavaré, S. and Ward, R. (1992). Modeling the evolution of the human mitochondrial genome. *Mathematical Biosciences* **112,** 319–335.

[30] Marth, G., Czabarka, E., Murvai, J. and Sherry, S. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166,** 351–372.

[31] McVean, G., Myers, S., Hunt, S., Deloukas, P., Bentley, D. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304,** 581–584.

[32] Möhle, M. and Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *The Annals of Probability* **29,** 1547–1562.

[33] Möhle, M. and Sagitov, S. (2003). Coalescent patterns in diploid exchangeable population models. *Journal of mathematical biology* **47,** 337–352.

[34] Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310,** 321–324.

[35] Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304.

[36] Nelson, M. R., Wegmann, D., Ehm, M. G. et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337,** 100–104.

[37] Pitman, J. (1992). The two-parameter generalization of Ewens' random partition structure. *Technical report* 345. Department of Statistics, U.C. Berkeley.

[38] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probabability Theory and Related Fields* **102,** 145–158.

[39] Polanski, A., Bobrowski, A. and Kimmel, M. (2003). A note on distributions of times to coalescence, under time-dependent population size. *Theoretical Population Biology* **63,** 33–40.

[40] Polanski, A. and Kimmel, M. (2003). New explicit expressions for relative frequencies of Single-Nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165,** 427–436.

[41] Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics* **78,** 629–644.

[42] Stephens, M. (2001). Inference under the coalescent. In *Handbook of Statistical Genetics*. ed. D. Balding, M. Bishop, and C. Cannings. Wiley, Chichester, UK pp. 213–238.

[43] Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics* **76,** 449–462.

[44] Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., GO, B., GO, S. and on behalf of the NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69.

[45] Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Roberts & Company Publishers.

[46] Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16,** 97–159.

[47] Wright, S. (1949). Adaptation and selection. In *Genetics, Paleontology and Evolution*. ed. G. L. Jepson, E. Mayr, and G. G. Simpson. Princeton University Press pp. 365–389.

[48] Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39,** 105–111.

# Appendix A

# Some combinatorial identities

Here, we provide some general combinatorial identities which are used several times for proving the main results in this thesis.

**Fact 1.** *For any positive integers $x, y, a$ and $b$ where $b \leq a$ and $x \leq y$,*

$$\sum_{m=x}^{y} \frac{\binom{b}{m}}{\binom{a}{m}} = \frac{\binom{a+1-x}{a+1-b} - \binom{a-y}{a+1-b}}{\binom{a}{b}}. \tag{A.1}$$

*Proof.* Starting from the left hand side of (A.1), we have:

$$\begin{aligned} \sum_{m=x}^{y} \frac{\binom{b}{m}}{\binom{a}{m}} &= \frac{b!(a-b)!}{a!} \sum_{m=x}^{y} \binom{a-m}{a-b} \\ &= \frac{\binom{a+1-x}{a+1-b} - \binom{a-y}{a+1-b}}{\binom{a}{b}}, \end{aligned}$$

where the last equality follows from the standard combinatorial identity that for all positive integers $a, n$, and $k$,

$$\sum_{i=a}^{n} \binom{n-i}{k} = \binom{n-a+1}{k+1}.$$

□

**Fact 2.** *For positive integers $a$ and $b$,*

$$\sum_{m=1}^{a} \frac{1}{m} \binom{a-m}{b} = \binom{a}{b} (H_a - H_b).$$

Fact 2 can be verified by induction [8] or by the method of Wilf-Zeilberger pairs [10].

**Fact 3.** *For positive integers $a$ and $b$ where $b \leq a$,*

$$\sum_{m=1}^{b} \frac{\binom{b}{m}}{\binom{a}{m}} \frac{1}{m+1} = \frac{a+1}{b+1} (H_{a+1} - H_{a-b}) - 1. \tag{A.2}$$

*Proof.* Starting from the left hand side of (A.2), we have:

$$\begin{aligned}
\sum_{m=1}^{b} \frac{\binom{b}{m}}{\binom{a}{m}} \frac{1}{m+1} &= \frac{b!(a-b)!}{a!} \sum_{m=1}^{b} \binom{a-m}{a-b} \frac{1}{m+1} \\
&= \frac{1}{\binom{a}{b}} \sum_{m=2}^{b+1} \binom{a+1-m}{a-b} \frac{1}{m} \\
&= \frac{1}{\binom{a}{b}} \left[ \sum_{m=1}^{b+1} \binom{a+1-m}{a-b} \frac{1}{m} - \binom{a}{b} \right] \\
&= \frac{1}{\binom{a}{b}} \left[ \binom{a+1}{b+1} (H_{a+1} - H_{a-b}) - \binom{a}{b} \right] \\
&= \frac{a+1}{b+1} (H_{a+1} - H_{a-b}) - 1,
\end{aligned}$$

where the fourth equality follows from using Fact 2. □

We also list some facts about the moments of a hypergeometric distribution which are appealed to several times in the thesis.

**Fact 4.** *If a multivariate hypergeometric distribution is parameterized by $\boldsymbol{n} = (n_1, \ldots, n_L)$, where $n = |\boldsymbol{n}|$, and a sample of size $m$, $\boldsymbol{m} = (m_1, m_2, \ldots, m_L)$, is drawn from it, then for any $\boldsymbol{t} = (t_1, t_2, \ldots, t_L)$ where $t_i \geq 0$ for all $i$, $t = |\boldsymbol{t}|$ and $t \leq n$,*

$$\mathbb{E}\left[ \prod_{i=1}^{L} (m_i)_{t_i\downarrow} \right] = \sum_{\substack{\boldsymbol{0} \preceq \boldsymbol{m} \preceq \boldsymbol{n}: \\ |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \prod_{i=1}^{L} (m_i)_{t_i\downarrow} = \frac{\prod_{i=1}^{L} (n_i)_{t_i\downarrow}}{(n)_{t\downarrow}} (m)_{t\downarrow} \tag{A.3}$$

*Proof.* Starting from the middle term in (A.3), we get:

$$\begin{aligned}
\sum_{\substack{\boldsymbol{0} \preceq \boldsymbol{m} \preceq \boldsymbol{n}: \\ |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} \prod_{i=1}^{L} (m_i)_{t_i\downarrow} &= \sum_{\substack{\boldsymbol{0} \preceq \boldsymbol{m} \preceq \boldsymbol{n}: \\ |\boldsymbol{m}| = m}} \frac{\prod_{i=1}^{L} (n_i)_{t_i\downarrow}}{(n)_{t\downarrow}} (m)_{t\downarrow} \frac{\binom{\boldsymbol{n}-\boldsymbol{t}}{\boldsymbol{m}-\boldsymbol{t}}}{\binom{n-t}{m-t}} \\
&= \frac{\prod_{i=1}^{L} (n_i)_{t_i\downarrow}}{(n)_{t\downarrow}} (m)_{t\downarrow} \sum_{\substack{\boldsymbol{0} \preceq \boldsymbol{m} \preceq \boldsymbol{n}-\boldsymbol{t}: \\ |\boldsymbol{m}| = m-t}} \frac{\binom{\boldsymbol{n}-\boldsymbol{t}}{\boldsymbol{m}}}{\binom{n-t}{m}} \\
&= \frac{\prod_{i=1}^{L} (n_i)_{t_i\downarrow}}{(n)_{t\downarrow}} (m)_{t\downarrow},
\end{aligned}$$

where the last equality follows because the term being summed is the probability mass function of a multivariate hypergeometric distribution parameterized by $\boldsymbol{n} - \boldsymbol{t}$, and the summation is over the entire domain of the distribution, and hence is 1. □

In the following fact, we compute some second moments of the hypergeometric distribution parameterized by $\boldsymbol{n}$ when restricted to those samples $\boldsymbol{m}$ which are non-zero at all types.

**Fact 5.** *If $\boldsymbol{n} = (n_1, n_2, \ldots, n_L)$, where $n = |\boldsymbol{n}|$, and $1 \le j \ne k \le L$, then we have the following identities:*

$$\sum_{\substack{\mathbf{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} m_j(m_j+1) = \sum_{\substack{T \subseteq [L]: \\ j \notin T}} (-1)^{|T|} \left[ \frac{(n_j)_{2\downarrow}(m)_{2\downarrow}}{(n-n_T)_{2\downarrow}} + \frac{2n_j m}{n-n_T} \right] \frac{\binom{n-n_T}{m}}{\binom{n}{m}} \tag{A.4}$$

$$\sum_{\substack{\mathbf{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} m_j m_k = \sum_{\substack{T \subseteq [L]: \\ j \notin T}} (-1)^{|T|} \frac{m_j m_k (m)_{2\downarrow}}{(n-n_T)_{2\downarrow}} \frac{\binom{n-n_T}{m}}{\binom{n}{m}} \tag{A.5}$$

*Proof.* Applying the inclusion-exclusion principle and using Fact 4, the identity in (A.4) can be obtained as

$$\begin{aligned}
\sum_{\substack{\mathbf{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} m_j(m_j+1) &= \sum_{\substack{T \subseteq [L]: \\ j \notin T}} (-1)^{|T|} \left[ \sum_{\substack{\mathbf{0} \preceq \boldsymbol{m} \preceq \boldsymbol{n} - \boldsymbol{n}_T: \\ |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}-\boldsymbol{n}_T}{\boldsymbol{m}}}{\binom{n-n_T}{m}} \left( (m_j)_{2\downarrow} + 2m_j \right) \right] \frac{\binom{n-n_T}{m}}{\binom{n}{m}} \\
&= \sum_{\substack{T \subseteq [L]: \\ j \notin T}} (-1)^{|T|} \left[ \frac{(n_j)_{2\downarrow}(m)_{2\downarrow}}{(n-n_T)_{2\downarrow}} + \frac{2n_j m}{n-n_T} \right] \frac{\binom{n-n_T}{m}}{\binom{n}{m}}.
\end{aligned}$$

Similarly for (A.5), we have

$$\begin{aligned}
\sum_{\substack{\mathbf{0} \prec \boldsymbol{m} \preceq \boldsymbol{n}: \\ |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}}{\boldsymbol{m}}}{\binom{n}{m}} m_j m_k &= \sum_{\substack{T \subseteq [L]: \\ j,k \notin T}} (-1)^{|T|} \left[ \sum_{\substack{\mathbf{0} \preceq \boldsymbol{m} \preceq \boldsymbol{n} - \boldsymbol{n}_T: \\ |\boldsymbol{m}| = m}} \frac{\binom{\boldsymbol{n}-\boldsymbol{n}_T}{\boldsymbol{m}}}{\binom{n-n_T}{m}} m_j m_k \right] \frac{\binom{n-n_T}{m}}{\binom{n}{m}} \\
&= \sum_{\substack{T \subseteq [L]: \\ j,k \notin T}} (-1)^{|T|} \frac{m_j m_k (m)_{2\downarrow}}{(n-n_T)_{2\downarrow}} \frac{\binom{n-n_T}{m}}{\binom{n}{m}}.
\end{aligned}$$

□