Frontiers in Coalescent Theory: Pedigrees, Identity-by-Descent, and Sequentially Markov Coalescent Models

Frontiers in Coalescent Theory: Pedigrees, Identity-by-descent, and Sequentially Markov Coalescent Models

a dissertation presented
by
Peter Richard Wilton
to
The Department of Organismic and Evolutionary Biology

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biology

Harvard University
Cambridge, Massachusetts
May 2016

Thesis advisor: Professor John Wakeley Peter Richard Wilton

Frontiers in Coalescent Theory: Pedigrees, Identity-by-descent, and Sequentially Markov Coalescent Models

Abstract

The coalescent is a stochastic process that describes the genetic ancestry of individuals sampled from a population. It is one of the main tools of theoretical population genetics and has been used as the basis of many sophisticated methods of inferring the demographic history of a population from a genetic sample. This dissertation is presented in four chapters, each developing coalescent theory to some degree. In the first chapter, I investigate how patterns of coalescence are affected by the population pedigree in structured populations, showing that the pedigree has longer-term effects in structured populations than in unstructured populations. Based on my findings, I develop a conceptual framework for jointly inferring population sizes, migration rates, and the recent pedigree of sampled individuals, and I demonstrate the efficacy of this approach in an application to simulated data. In Chapter 2, I present a theoretical study of the distribution of segments of identity-by-descent, showing how the accuracy of predictions made based on sequentially Markov coalescent models depends on the particular model being used as the basis of calculations. In the third chapter, I undertake a theoretical comparison of two approximations, termed the SMC and SMC', to the full model of coalescence with recombination. I derive new theoretical properties of the SMC' and use these properties to demonstrate that the SMC' is, in a well-defined sense, the most appropriate first-order approximation to the full coalescent with recombination. I also show that estimates of population size based on the SMC are statistically inconsistent. Finally, in Chapter 4, I develop a coalescent hidden Markov model approach to inferring the demographic and reproductive history of a triploid asexual lineage derived from a diploid sexual ancestor. The motivation for this project is an ongoing collaborative effort to sequence and analyze the genomes of sexual and asexual lineages of the New Zealand snail *Potamopyrgus antipodarum*. The method I present in this chapter will be applied to these genomes to infer when triploid asexual lineages were derived from sexual ancestors and to describe the demographic history of those sexual ancestors. Here, I investigate the this method with simulated asexual genomes.

Contents

Listing of figures

Dedicated to Shirley Wilton, Debra Fincham, and Richard Fincham.

Acknowledgments

There are several people I need to thank who provided help and support during my time as a graduate student. First, my advisor, John Wakeley, has given me consistently thoughtful advice, frequent encouragement, and the freedom to pursue the research topics that interest me, and I am tremendously grateful to him. My discussions with John have been a large part of my education as a scientist. I also thank the members of my committee, Michael Desai, Scott Edwards, and Jim Mallet, for the guidance they have provided.

Much of the work I have done as a graduate student has been in collaboration with other researchers. I am thankful to Pierre Baduel and Matthieu Landon for the contributions they made to the work on pedigrees presented in Chapter 1. Shai Carmi, Pier Palamara, and Asger Hobolth were instrumental to the work I present on identity-by-descent and the SMC' in Chapters 2 and 3. Maurine Neiman and the rest of the *Potamopyrgus* genome team have improved the demographic inference method presented in Chapter 4 by providing much-needed empirical perspective. I am especially grateful to my friend Frank Rheindt, with whom I have been lucky enough to collaborate on multiple projects involving birds, which I have enjoyed thoroughly.

My graduate school experience would not have been nearly as enjoyable as it was without the friends with whom I have shared my life the last several years. I am tremendously thankful for the support and companionship I have received from Seth Donoughe, Christina Baik, Dan Rice, Mary Schnoor, Kara Feilich, Brent Hawkins, Ambika Kamath, Joel Nitta, José Rojas, Katie Boronow, Maria Metzler, Bryan McCullough, Léandra King, Julia Palacios, Joel Nitta, Wenfei Tong, Frank Rheindt, and many others.

I am lucky to have a loving and supporting family. My mother and father Janice and Thomas Wilton have always encouraged my curiosity and my desire for learning. I also thank my brother, Brian Wilton, for the love and friendship he shares with me.

Finally, I thank my fiancée Jennifer Holle, who fills my life with joy.

Introduction

We are living in an era of abundant information. From the devices we carry to the purchases we make and the media we consume, seemingly every interaction we have with the world generates information that is stored and analyzed for one purpose or another. We also carry information in our cells, written in the three billion DNA base pairs that make up a human genome. The particular sequence of these base pairs encodes nearly all of the information that is required for life, and it determines much of what makes each of us the individuals we are. Scientists first discovered how to sequence DNA molecules in the 1970's, and since then we have been amassing genetic information in the form of DNA sequences at a rapidly accelerating pace. The human genome sequence was published in 2003 after more than a decade of collaboration between researchers at numerous institutions. With today's sequencing technology, it takes about a day to produce a human genome of moderate quality, and this time will soon shorten as new sequencing technologies become available. There are currently thousands of publicly available human genome sequences, and it was recently estimated that altogether the DNA sequencing machines currently in use have the capacity to produce about 35 petabases of sequence per year [1]. As this DNA sequence data continues to accumulate, it becomes imperative that we have the tools to efficiently extract the information contained in the sequences.

The promise of DNA sequencing has been that it will provide medical breakthroughs and pave the way for precision medicine, where medical treatment is tailored to the patient's genome. DNA sequencing has transformed biomedical research — many of the promised breakthroughs have been provided, and personalized, precision medicine still drives diverse research efforts and enormous investments.

Perhaps a less obvious use of DNA sequencing is to use it to learn about our collective evolutionary history. As DNA is transmitted from parent to offspring, occasional errors in the DNA replication process cause changes in the DNA sequence to be inherited by the offspring. These mutations may subsequently be inherited by individuals in future generations, leaving a signature of the reproductive success of the ancestors of any present-day individual bearing that mutation. When many individuals and mutations are considered, patterns of genetic variation emerge in the DNA sequence data. These patterns depend on the dynamics of reproduction in the population: What is the size of the population? How is reproduction structured across geographic space? Do some individuals have a heritable advantage over others, leading to natural selection? How has all this changed over time? Change the answer to any of these questions, and patterns of genetic variation in the present-day population will also change.

To determine how the dynamics of reproduction affect the distribution of genetic variation in a population, it is necessary to construct a mathematical model of reproduction. This is the domain of the field of theoretical population genetics, and it is, broadly, the subject of this dissertation. The foundations of theoretical population genetics were established in the early and middle twentieth century, before the advent of DNA sequencing. Classic work by Haldane [2], Wright [3], Fisher [4], Kimura [5], and others showed how random mating, mutation, natural selection, migration, and other forces acting in a population conspire to shape patterns of genetic variation in the population. Many of the models developed by these authors are prospective: Given that a variant is at a particular frequency in the present generation, the models studied by these authors predict the distribution of trajectories that the frequency might take over the course of future generations (Fig. 1A). This forward-looking “diffusion” approach to modeling populations is especially well suited for gaining a conceptual understanding of how evolution works, since evolution happens forward in time.

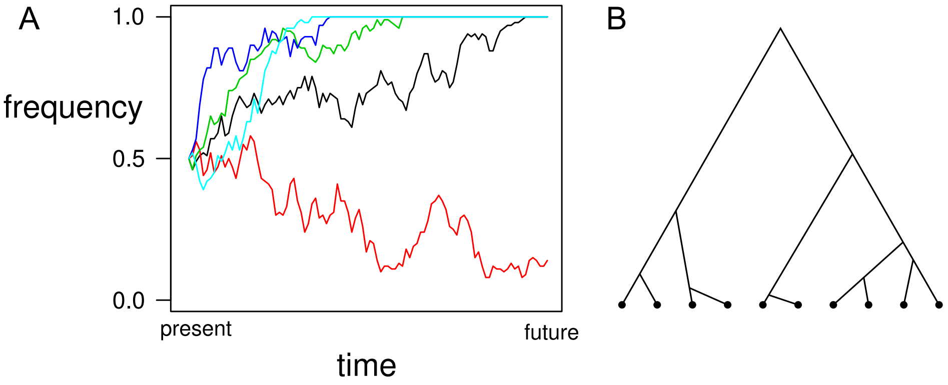

Figure 1: Illustration of forward-in-time and backward-in-time approaches to theoretical population genetics. In panel **A**, five different trajectories are shown for a genetic variant initially at frequency 0.5 in the population. Random mating by individuals in the population causes the frequency of the variant to change over time. Diffusion theory makes probabilistic statements about the distribution of these trajectories. In panel **B**, a gene genealogy for a sample of ten copies of a gene is shown. This tree-like structure shows the relationships of the ancestors of the sampled gene copies. Each merging of two branches represents a “coalescent event,” where some single ancestor had two offspring that each gave rise to a separate lineage that is also ancestral to the sample. Coalescent theory makes predictions about gene genealogies and uses these structures to make predictions about genetic variation in samples. Diffusion models and coalescent models are closely linked to one another, and the same results can often be arrived at using both approaches.

Another, more recently introduced approach to modeling genetic variation in a population is retrospective. Given a sample of genes, this approach models the ancestry of the sample as a stochastic process. At different points in the past, pairs of genetic lineages ancestral to the sample reach their common ancestor in generations where one ancestor had two offspring that were each themselves ancestral to the sampled gene copies. Eventually a common ancestor of the entire sample is reached, and the resulting tree structure, termed a “gene genealogy,” fully describes the ancestry of the sample (see Fig. 1B). The model for generating these gene genealogies is called the coalescent [6], and the subfield of theoretical population genetics that deals with gene genealogies is called coalescent theory. By allowing mutations to occur along the branches of a gene genealogy, it is possible to directly model genetic variation in a sample. This makes coalescent theory a powerful and flexible framework for inferring the demographic and evolutionary history of a population.

This dissertation presents my graduate school work in four chapters, each involving and developing coalescent theory to some degree. The first chapter addresses the question of how the biparental pedigree, *i.e.* genealogy, of a population affects the process of coalescence. For mathematical convenience, in population genetics it is typically assumed that each chromosome is inherited through a different pedigree, which is not actually the case. This chapter grows out of recent work showing that the particular shape of the pedigree has minimal effects on coalescence in random-mating, completely unstructured populations [7]. Here, I consider how the pedigree affects coalescent patterns in structured populations, where multiple subpopulations are connected by occasional migration events between them. I show that compared to the case of unstructured populations, the pedigree has larger effects over a longer period than in the unstructured case. This work also produced a new conceptual approach for jointly inferring population sizes, migration rates, and the pedigree of the sample. I conclude Chapter 1 with a demonstration of this approach, inferring these parameters from simulated data.

In Chapter 2, I explore the connections between coalescent theory and identity-by-descent, which occurs when two individuals both inherit a stretch of their genome from the same, recent ancestor. I provide a technical, theoretical update to previous work modeling identity-by-descent using coalescent theory, performing calculations based on a more complex but more accurate model of coalescence and recombination, termed the SMC'. This part of Chapter 2 was incorporated into in a paper I co-authored with Shai Carmi, my advisor John Wakeley, and Itsik Pe'er [8]. The second half of this chapter deals with making quantitative predictions about mutation along stretches of identity-by-descent. This work was part of a paper I co-authored with Pier Palamara and others [9].

Chapter 3 explores theoretical properties of models of coalescence and recombination. The full idealized model of recombination and coalescence, called the "ancestral recombination graph" or sometimes "coalescent with recombination," is a relatively simple model to describe, but many mathematical obstacles lay in the way of its use in coalescent-based statistical inference. Two approximations to the ancestral recombination graph, the original sequentially Markov coalescent (SMC) [10] and then another called the SMC' [11], were introduced as tractable approximations to the full ancestral recombination graph. These models have been used as a basis for many of the most recent sophisticated population-genetic inference procedures [*e.g.,* 12, 13, 14]. In Chapter 3, I provide a variety of new theoretical results for the SMC' model of coalescence and recombination. I use these results to show that the SMC' is, in a certain sense, the most suitable approximation to the full model. I also show that inferences made based on the simpler SMC model are statistically inconsistent, *i.e.* they will produce biased estimates even with infinite amounts of data. This work was done in collaboration with Shai Carmi and Asger Hobolth, with whom I co-authored a paper presenting the work in this chapter [15].

Finally, in Chapter 4, I present a demographic inference method tailored to triploid asexual lineages of the New Zealand snail *Potamopyrgus antipodarum.* The method uses the genome of such an asexual snail to infer the time at which a triploid snail lineage transitioned to asexual reproduction from sexual ancestors, as well as the population size history of those sexual ancestors. Like other recent coalescent-based demographic inference methods [12, 13, 16, 17], the inference procedure I propose uses a hidden Markov model derived from a sequentially Markov coalescent model (here, the SMC'). The innovation I provide is to fully model the genealogies of more than two sequences under the standard coalescent. This is made possible by averaging over the phasing of the three chromosomes in the triploid asexual genome. This project is motivated by a collaboration with Maurine Neiman and others, who are sequencing the genomes of many sexual and asexual *P. antipodarum*. Currently, I have completed the theoretical phases of this project; once the sequencing is completed later this year, I will apply the inference procedure described in this chapter to the sequenced genomes. In the meantime, in this dissertation I describe the method and explore its accuracy when applied to simulated data, and I discuss some of its potential limitations.

1

Population structure and coalescence in pedigrees

1.1 Introduction

The coalescent is a stochastic process that describes how to construct gene genealogies, the tree-like structures that relate the sampled copies of a gene to one another. Since its introduction by Kingman [6, 18], the coalescent has been extended and applied to numerous areas in population genetics and is now one of the foremost mathematical tools for modeling genetic variation [19, 20].

In a typical application to data from diploid sexual organisms, the coalescent is applied to multiple loci that are assumed to have independent ancestries because they are found on different chromosomes and thus segregate independently, or are far enough apart along a single chromosome that effectively they segregate independently. Even chromosome-scale coalescent-based inference methods [e.g., 12, 13, 17] multiply probabilities across distinct chromosomes that are assumed to have completely independent histories due to their independent segregation.

In a diploid sexual population, all genetic material is inherited through a single population pedigree. The population pedigree is the structure that contains all relationships between all members of the population throughout all time. In most populations the pedigree is unobserved, but in cases where the pedigree has been entirely or partially observed, either through field observation, thorough genetic sampling, or by examining historical records (in the case of humans), it can be highly informative about past and present demographic and evolutionary forces in the population [21, 22]. Regardless of whether the pedigree has been observed, it is true that each sexual diploid population has only one pedigree, and all all of the sampled chromosomes segregated through that same pedigree. In assuming that independently segregating loci have completely independent gene genealogies, it is implicitly also assumed that each such locus was inherited through an independent population pedigree.

While this is clearly not the case, the non-independence between gene genealogies of independently segregating loci introduced by the shared population pedigree has only recently been examined. Wakeley et al. [7] studied gene genealogies of loci segregating independently through pedigrees of diploid populations generated under basic Wright-Fisher-like reproductive dynamics, i.e., populations with constant population size, random mating, non-overlapping generations, and lacking population structure. In this context, they found that the shape of the population pedigree affected coalescence probabilities mostly in the first $\sim \log_2(N)$ generations back in time, where N is the population size, and they found that in general it was difficult to distinguish distributions of coalescence times that were generated by segregating independent chromosomes back in time through a fixed, randomly-generated pedigree from the predictions of the standard coalescent.

That the pedigree should have substantial effects on coalescence probabilities only during the most recent $\sim \log_2(N)$ generations is in agreement with other theoretical studies of the structure of population pedigrees. Chang [23] found that the number of generations until two individuals share an ancestor in the biparental, pedigree sense converges to $\log_2(N)$ as the population size grows. Likewise, Derrida et al. [24] showed that the distribution of the number of repetitions in an individual's pedigree ancestry becomes stationary around $\log_2(N)$ generations in the past. This $\log_2(N)$-generation timescale is the natural timescale of convergence in pedigrees due to the approximate doubling of the number of possible ancestors each generation back in time until the ancestral population size is reached. This occurs around $\sim \log_2(N)$ generations in the past in a well mixed, constant-sized population of size N.

In each of these studies it is assumed that the population is panmictic, i.e., that individuals mate with each other uniformly at random. One phenomenon that may alter this convergence in pedigrees is population structure, with migration between subpopulations. In a subdivided population, the exchange of ancestry between subpopulations depends on the particular history of migration events embedded in the population pedigree. These past migration events may be infrequent or irregular enough that the convergence in the pedigree depends on the details of the migration history rather than on the reproductive dynamics underlying convergence in unstructured populations. Rohde et al. [25] studied the sharing of pedigree ancestry in structured populations and found that population structure did not change the $\log_2(N)$-scaling of the number of generations until a common ancestor of everyone in the population is reached. Barton and Etheridge [26] studied the expected number of descendants of an ancestral individual, a quantity termed the reproductive value, and similarly found that population subdivision did not much slow the convergence of this quantity over the course of generations.

While these results give a general characterization of how pedigrees are affected by population structure, a direct examination of the coalescent process for loci segregating independently through a fixed pedigree of a structured population is still needed. It may be that fixing the migration events in the pedigree produces long-term fluctuations in coalescent probabilities that make the predictions of the structured coalescent break down. Here, we explore how population structure affects coalescence through a fixed population pedigree. Using simulations, we investigate the fluctuations in coalescence probabilities caused by the variation in the migration history embedded in the pedigree and determine how these fluctuations depend on deme size and migration rate. We also study the particular effects of recent admixture on coalescence-time distributions and use our findings to develop a simple framework for modeling the ancestry of the sample as a mixture of different non-admixed ancestries. To demonstrate the efficacy of this approach, we create a maximum-likelihood inference procedure for inferring population-scaled migration and mutation rates jointly with the recent pedigree of the sample. Finally, we perform simulations to confirm that this approach allows for accurate inference of migration rates even in the presence of recent admixture in the sample.

1.2 Theory and Results

1.2.1 Pedigree simulation

Except where otherwise stated, each population we model has two sub-populations of constant size, exchanging migrants symmetrically at a constant rate. This model is simple to describe, demonstrates the effects of population structure in one of the the simplest ways possible, and has a relatively simple mathematical theory [20]. We anticipate that many of our results will be generalizable to more complex models of population structure.

We assume that generations are non-overlapping and that the population has two sexes that are equal in number. In each generation, each individual chooses a mother uniformly at random from the females of the same deme with probability $1 - m$ and from the females of the other deme with probability m. Likewise a father is chosen uniformly at random from the males of the same deme with probability $1 - m$ and from the males of the other deme with probability m. This particular model of migration corresponds to broadcast spawning, in which gametes migrate but individuals do not. Through simulation of other models of reproduction and migration, we find that our results are not sensitive to these particulars of the migration process.

All simulations were carried out with `coalseam`, a program for coalescent simulation through randomly-generated population pedigrees. The user provides parameters such as population size, number of demes, mutation rate, and migration rate, and `coalseam` simulates a population pedigree under a Wright-Fisher-like model meeting the specified conditions. Gene genealogies are constructed by simulating segregation back in time through the pedigree, and the resulting genealogies are used to produce simulated genetic loci. Output is in a format similar to that of the program `ms` [27], and various

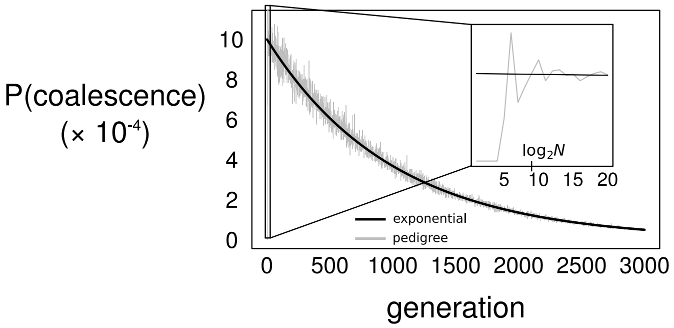

Figure 1.1: Coalescence time distribution for independently segregating loci sampled from two individuals in a panmictic population. The gray shows the distribution from the pedigree, and the black line shows the exponential prediction of the standard coalescent.

options allow the user to simulate and analyze pedigrees featuring, for example, recent selective sweeps or fixed amounts of identity by descent and admixture.

The program `coalseam` is written in C and released under a permissive license. It is available online at `https://github.com/ammodramus/coalseam`.

1.2.2 Structured population pedigrees and probabilities of coalescence

In a well-mixed population of size N, the distribution of coalescence times for independently segregating loci sampled from two individuals shows large fluctuations over the first $\sim \log_2(N)$ generations depending on the degree of overlap in the pedigree of the two individuals. After this initial period, the coalescence probabilities quickly converge to the expectation under standard coalescent theory, with small fluctuations around that expectation. [7, Fig. 1.1]. The magnitude of these fluctuations depends on the population size, but even for small to moderately sized populations (e.g., $N = 500$), the exponential prediction of the standard coalescent is a good approximation to the true distribution after the first $\log_2(N)$ generations.

When the population is divided into multiple demes, deviations from the coalescent probabilities predicted by the structured coalescent depend on the particular history of migration in the population pedigree. The effect of the migration history is especially pronounced when the average number of migration events per generation is of the same order as the per-generation pairwise coalescent probability (Fig. 1.2A). In this migration-limited regime, two lineages in different demes have zero probability of coalescing before a migration event in the pedigree can bring them together into the same deme. This creates large peaks in the coalescence time distributions for loci segregating independently through the same pedigree, with each peak corresponding to a particular migration event (Fig. 1.2A). This scenario is somewhat implausible, however: It will not often be the case that migration is so infrequent that it happens on a population-wide level only every $\sim 1/N$ generations. In such a scenario, it would be more realistic to model the migration process as a series of distinct admixture pulses rather than migration occurring at a continuous rate, even if (as here) the underlying migration process does have a constant rate.

Even when the migration rate is higher and there are many migration events per coalescent event, the pedigree can still cause coalescence probabilities to differ from the predictions of the structured coalescent. Under these conditions, coalescence is not constrained by individual migration events, but there may be stochastic fluctuations in the realized migration rate, with some periods experiencing more migration and others less. These fluctuations can cause deviations in the predicted coalescence probabilities long past the $\log_2(N)$-generation timescale found in well-mixed populations (Figs. 1.2B, S1, S2). The degree of these deviations depends on the migration rate and the population size, with smaller populations and lower migration rates causing greater deviations, and deviations from predictions are generally larger for samples between demes than sam-

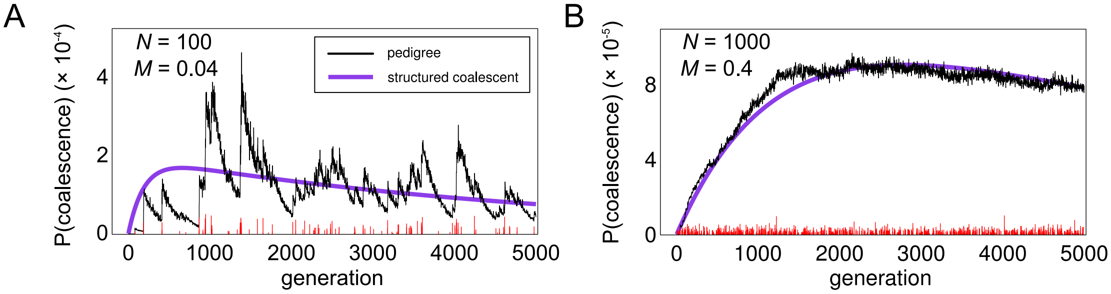

Figure 1.2: Distribution of pairwise coalescence times for two individuals sampled from different demes. In both panels, the black line shows the distribution calculated from the simulated pedigrees, and the purple line shows the prediction from the structured coalescent. Vertical lines along the horizontal axis show the occurrence of migration events in the population, with the relative height representing the total reproductive weight [see 26] of the migrant individual(s) in that generation. **(A)** Low migration pedigree, with $M = 4Nm = 0.04$ and $N = 100$. Under these conditions, coalescence is limited by migration events, so there are distinct peaks in the coalescence time distribution corresponding to individual migration events. **(B)** Higher migration pedigree, with $M = 0.4$ and $N = 1000$. With the higher migration rate, coalescence is no longer limited by migration, but stochastic fluctuations in the migration process over time cause deviations away from the standard-coalescent prediction on a timescale longer than $\log_2(N)$ generations.

ples within demes. Reassuringly, when there are many migration events per coalescent event (i.e., when $Nm >> 1/N$), the predictions of the structured coalescent seem to fit the observed distributions in pedigrees reasonably well (Figs. S1–S2).

To investigate the dependence of the coalescence time distribution on the pedigree more systematically, we simulated 20 replicate population pedigrees for a range of populations sizes and migration rates. From each pedigree, we sampled two individuals in different demes and calculated the distribution of pairwise coalescence times for independently segregating loci sampled from those two individuals. We measured the total variation distance from the distribution predicted under a discrete-time model of coalescence and migration analogous to the continuous-time structured coalescent. The total variation distance of two discrete distributions P and Q is defined as

$$D_{TV}(P,Q) = \frac{1}{2}\sum_{i} |P(i) - Q(i)| \, . \tag{1.1}$$

We found that the total variation distance between the distributions of pairwise coalescence times from pedigrees and the distributions from standard theory decreases as both N and M increase and that, in general, the total variation distance is more sensitive to the migration rate than on population size (Fig. S3).

1.2.3 Admixture and coalescence distributions in pedigrees

As is the case for panmictic populations, the details of the *recent* sample pedigree are most important in determining the patterns of genetic variation in the sample. In panmictic populations, overlap in ancestry in the recent past creates identity-by-descent. In structured populations, an individual may also have recent relatives from another deme, resulting in admixed ancestry. When this occurs, the distribution of pairwise coalescence times is potentially very different from the prediction in the absence of admixture due to the admixture paths in the pedigree that lead to a recent change in demes. The degree of the difference in distributions is directly related to the degree of admixture, with more recent admixture causing greater changes in the coalescence time distribution (Fig. 1.3).

The distribution of coalescence times for a sample with admixed ancestry can be approximated by considering the sample to be a mixture of different samples. With admixture in the recent sample pedigree, there is some probability that after the first few generations back in time, one or more of the lineages will not be in the deme from which they were originally sampled. When this happens, the genetic variation in the

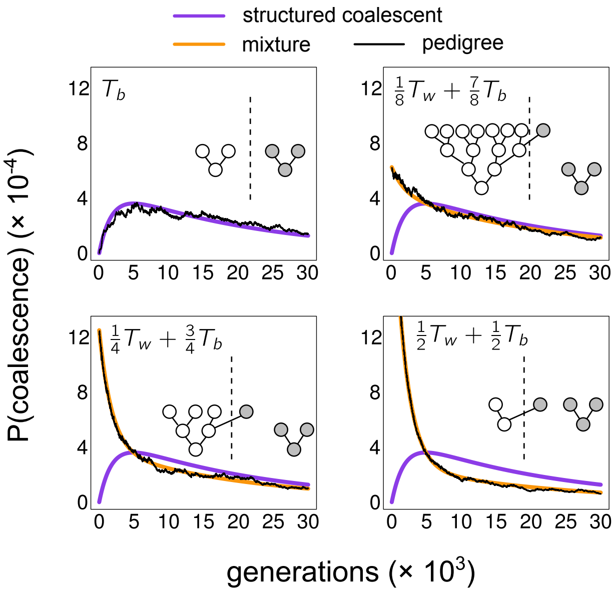

Figure 1.3: Pairwise coalescence time distributions for samples with admixed ancestry. Each panel shows the distribution of pairwise coalescence times for a sample whose pedigree contains some amount of admixture. In each simulation, there are two demes of size $N = 1000$, and the scaled migration rate is $4Nm = 0.1$. In each panel, the population pedigree was simulated conditional on the sample having the pedigree shown in the panel. The purple line shows the between-deme coalescence time distribution that would be expected in the absence of admixture, and the gold line shows the the mixture of the within- and between-deme coalescence time distributions that corresponds to the degree of admixture. Black lines are numerically calculated coalescence time distributions for the simulated example pedigrees.

sample will reflect the locations of the lineages before the admixture took place. This type of "sample reconfiguration" can be viewed probabilistically, with the Mendelian probabilities of the different paths through the pedigree determining the probabilities of different sample reconfigurations.

As an example, consider a sample of unlinked loci taken from two individuals related by the pedigree shown in Fig. 1.3C, where one of two individuals sampled from different demes has a grandparent from the other deme. The distribution of pairwise coalescence times for loci sampled from this pair resembles the distribution of $T_w/4 + 3T_b/4$, where T_b is the standard between-deme pairwise coalescence time for a structured-coalescent model with two demes, and T_w is the corresponding within-deme pairwise coalescence time (Fig. 1.3C). The particular mixture reflects the fact that a lineage sampled from the admixed individual follows the admixture path with probability 1/4.

This sample reconfiguration framework can also be used to model identity-by-descent (IBD), where overlap among branches of the recent sample pedigree causes early coalescence with unusually high probability. If the pedigree causes an IBD event to occur with probability Pr(IBD), then the pairwise coalescence time is a mixture of the standard distribution (without IBD) and instantaneous coalescence (on the coalescent timescale) with probabilities $1 - \Pr(\text{IBD})$ and Pr(IBD), respectively. If there is both IBD and admixture in the recent sample pedigree (or if there are multiple admixture or IBD events), the sample can be modeled as a mixture of several sample reconfigurations (e.g., fig. 1.4).

This approach to modeling the sample implicitly assumes that there is some threshold generation separating the recent pedigree, which determines the mixture of sample reconfigurations, and the more ancient pedigree, where the standard coalescent models are assumed to hold well enough. The natural boundary between these two periods is

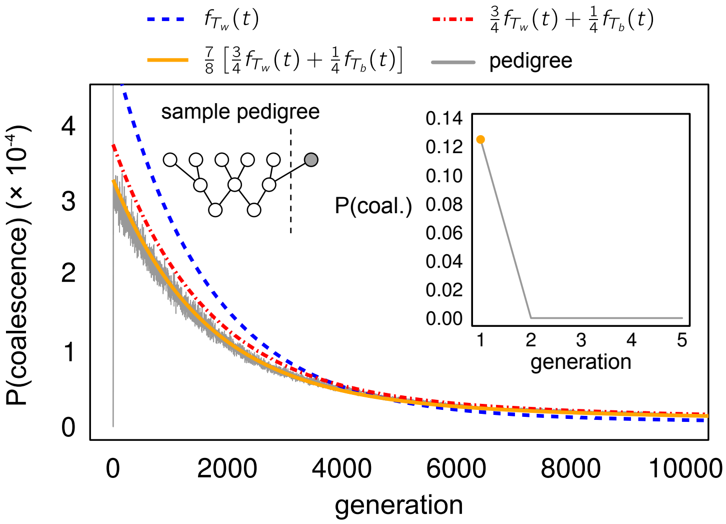

Figure 1.4: Distribution of pairwise coalescence times for a sample whose recent pedigree contains both admixture and IBD. The recent pedigree of the sample is shown, with the two sampled individuals located at the bottom of the pedigree. The distribution for the simulated pedigree (gray line) is based on numerical coalescence probabilities calculated in a pedigree of two demes of size $N = 1000$ each, with migration rate $M = 4Nm = 0.2$. The colored lines show mixtures of the within-deme coalescence time distribution (f_{T_w}) and the between-deme distribution (f_{T_b}). The inset shows the probability of coalescence during the first five generations; the probability mass at generation 1 is predicted by the mixture model accounting for both IBD and admixture (orange line). The red line shows the distribution if IBD is ignored, and the blue line shows the distribution if both IBD and admixture are ignored.

around $\log_2(N)$ generations, since any pedigree feature more ancient than that tends to be shared by most or all of the population (making such features "population demography"), and any features more recent tend to be particular to the sample. In practice, it seems sufficient to model only the first $\sim$ 3–5 generations back in time, since events beyond this time have relatively minor effects on patterns of coalescence.

We note that there is a long history in population genetics of modeling genetic variation in pedigrees as a mixture of different sample reconfigurations. Wright [28] wrote the probability of observing a homozygous A_1A_1 genotype as $p^2(1-F)+pF$, where p is the frequency of A_1 in the population and F is essentially the probability of IBD calculated from the sample pedigree. This can be thought of as a probability for a mixture of two samples of size $n = 2$ (with probability $1-F$) and $n = 1$ (probability F). The popular ancestry inference program STRUCTURE [29] and related methods similarly write the likelihood of observed genotypes as a mixture over different possible subpopulation origins of the sampled alleles. Here, motivated by our simulations, we explicitly extend this approach to coalescent models. In the next section, we create an example method for inferring population parameters such as mutation and migration rates jointly with features of the sample pedigree such as recent IBD and admixture. In the Discussion (see below), we further discuss the similarities and differences between our modeling approach and those of existing methods.

1.2.4 Joint inference of the recent sample pedigree and population demography

The sample reconfiguration framework for modeling genetic variation in pedigrees shows how estimators of population-genetic parameters may be biased by admixture or IBD in the recent sample pedigree. In Appendix A, we calculate the bias of three estimators of the population-scaled mutation rate $\theta = 4N\mu$ in a panmictic population when the recent sample pedigree contains IBD. In Appendix B, we calculate the bias of a moments-based estimator of M due to recent admixture in the sample pedigree. We use simulations to confirm these calculations (Figs. S5,S6). In both cases, if the recent sample pedigree is known, it is straightforward to correct these estimators to eliminate the bias.

It is uncommon that the recent pedigree of the sample is known, however, and if it is assumed known, it is often estimated from the same data that is used to infer demographic parameters. Ideally, one would infer long-term demographic history jointly with sample-specific features of the pedigree. In this section, we develop a maximum-likelihood method for inferring IBD and admixture jointly with scaled mutation and migration rates. The method uses the approach proposed in the previous section: the sample pedigree defines some set of possible outcomes of Mendelian segregation in recent generations (ending approximately $\log_2(N)$ generations ago), and the resulting, reconfigured sample is modeled by the standard coalescent process.

Before we describe our method, we define some notation. We study a population with two demes each of size N, with each individual having probability m of migrating to the other deme in each generation. We rescale time by N so that the rate of coalescence within a deme is unity and the rescaled rate of migration per lineage is $M/2 = 2Nm$. We assume that we have sampled two copies of each locus from each of n_1 (diploid) individuals from deme 1 and n_2 individuals from deme 2. We write the total diploid sample size as $n_1 + n_2 = n$ so that the total number of sequences sampled at each locus is $2n$.

We index our sequences with $\mathcal{I}_n := \{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}, 2^{\mathrm{p}}, \dots, n^{\mathrm{m}}, n^{\mathrm{p}}\}$, where i^{m} and i^{p} index the maternal and paternal sequences sampled from individual i. Arbitrarily, we assume that the indices pertaining to the first n_1 individuals index sequences sampled from deme 1 and indices pertaining to the last n_2 individuals index the sequences sampled from deme 2.

Each recent pedigree $\mathcal{P}$ has some set of possible outcomes of segregation, involving coalescence of lineages (IBD) and movement of lineages between demes (admixture). The set of these reconfigurations is denoted $\mathcal{R}(\mathcal{P})$, and each reconfiguration $r \in \mathcal{R}(\mathcal{P})$ is a partition of $\mathcal{I}_n$, with the groups in r representing the lineages that survive after segregation back in time through the recent pedigree. Each group in a reconfiguration is also labeled with the deme in which the corresponding lineage is found after segregation back in time through the recent pedigree. The pedigree also induces a probability distribution $\Pr(r \mid \mathcal{P})$, $r \in \mathcal{R}(\mathcal{P})$, giving the probabilities of the different sample reconfigurations.

The data $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_l\}$ consist of sequence data at L loci. The data at locus i are represented by the sequences $\boldsymbol{X}_i = \{X_{i,1}^{(\mathrm{a})}, X_{i,1}^{(\mathrm{b})}, X_{i,2}^{(\mathrm{a})}, X_{i,2}^{(\mathrm{b})}, \ldots, X_{i,n}^{(\mathrm{a})}, X_{i,n}^{(\mathrm{b})}\}$, where $X_{i,j}^{(\mathrm{a})}$ and $X_{i,j}^{(\mathrm{b})}$ are the two sequences at locus i from individual j. We label them (a) and (b) because we assume that they are of unknown parental origin. In order to make calculation of sampling probabilities feasible, we assume that each sequence evolves under the infinite-sites mutation model and can thus be represented by a binary sequence. We also assume that there is free recombination between loci and no recombination within loci.

Our goal is to find the θ, M, and $\mathcal{P}$ that maximize the likelihood

$$\begin{aligned} L(\mathcal{P}, \theta, M \mid \boldsymbol{X}) = \Pr(\boldsymbol{X} \mid \mathcal{P}; \theta, M) = \\ \prod_{i=1}^{L} \sum_{r \in \mathcal{R}(\mathcal{P})} \Pr(\boldsymbol{X}_i \mid r; \theta, M) \Pr(r \mid \mathcal{P}). \end{aligned} \tag{1.2}$$

Probabilities are multiplied across independently segregating loci because their ancestries are independent conditional on the pedigree.

In order to calculate $\Pr(\boldsymbol{X}_i \mid r; \theta, M)$, it is necessary to consider all the ways the sequences $\boldsymbol{X}_i$ could have been inherited maternally versus paternally, since we assume that we do not know which sequence is maternal and which is paternal. For sequences $\boldsymbol{X}_i$, let $\Lambda(\boldsymbol{X}_i)$ be the set of all possible ways of labeling $\boldsymbol{X}_i$ as maternal and paternal, thus associating each sequence with an index in $\mathcal{I}_n$. The sampling probability of the data at locus i is then

$$\Pr(\boldsymbol{X}_i \mid r; \theta, M) = \frac{1}{2^n} \sum_{\lambda \in \Lambda(\boldsymbol{X}_i)} \Pr(X_i \mid \lambda, r; \theta, M), \tag{1.3}$$

since there are 2^n ways that the X_i could have segregated as maternal and paternal alleles, and each is equally likely to have occurred.

Together the reconfiguration $r \in \mathcal{R}(\mathcal{P})$ and the maternal-paternal labeling $\lambda \in \Lambda(\boldsymbol{X}_i)$ imply a partition $\mathbb{P}(\boldsymbol{X}_i, r, \lambda)$ of the sequences $\boldsymbol{X}_i$ corresponding to the partition of sequence indices represented by r. For each group $h \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda)$, there is a group $g \in r$ that can be mapped onto h such that 1) each index $i \in g$ indexes a distinct sequence in h sampled from the individual indexed by i, and 2) the deme labeling of g matches the deme labeling of h. Denote the unique elements of the set A as $A_{\neq}$. The conditional sampling probability of the sequences $\boldsymbol{X}_i$ given maternal-paternal labeling $\lambda \in \Lambda(\boldsymbol{X}_i)$ and sample reconfiguration $r \in \mathcal{R}(\mathcal{P})$ is

$$\begin{aligned} &\Pr(X_i \mid \lambda, r; \theta, M) = \Pr(\mathbb{P}(\boldsymbol{X}_i, r, \lambda); \theta, M) = \\ &\begin{cases} \phi(\{h_{\neq} : h \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda)\}; \theta, M) & \text{if } |h_{\neq}| = 1 \ \forall h \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda) \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{1.4}$$

where each $h \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda)$ is one of the non-empty subsets in the partitioned sequences, $|h_{\neq}|$ is the number of unique elements in such a subset, $\{h_{\neq} : h \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda)\}$ is the "reduced" set of sequences, such that each subset in the partition is replaced by the unique elements in the subset, and $\phi(\{h_{\neq} : g \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda)\}; \theta, M)$ is the standard infinite-sites sampling probability of the sample after it has been reconfigured by the recent pedigree. This sampling probability can be calculated numerically using a dynamic programming approach [30, 31, see below].

In other words, conditional on certain sequences being IBD (i.e., they are in the same group in the partitioned sequences), the sampling probability is the standard infinite-sites probability of the set of sequences with duplicate IBD sequences removed and the deme labelings of the different groups made to match any admixture events that may have occurred. If any of the sequences designated as IBD are not in fact identical in sequence, the sampling probability for that reconfiguration and maternal-paternal labeling is zero. (This assumes that no mutation occurs in the recent part of the pedigree.) Conveniently, each reconfigured sample with non-zero probability corresponds to one of the ancestral sample configurations in the recursion to solve the standard infinite-sites sampling probability for the *entire* sample $\boldsymbol{X}_i$, so that the sampling probability for all pedigrees can be calculated by solving the recursion for the sampling probability of the original sample only once.

Together, (1.2), (1.3), and (1.4) give the overall joint log-likelihood of mutation rate θ, migration rate M, and pedigree $\mathcal{P}$ given sequences $\boldsymbol{X}$:

$$
\begin{aligned}
&LL(\theta, M, \mathcal{P} \mid \boldsymbol{X}) = \\
&\sum_{i=1}^{L} \log \left(\sum_{r \in \mathcal{R}(\mathcal{P})} \Pr(r \mid \mathcal{P}) \sum_{\lambda \in \Lambda(\boldsymbol{X}_i, r)} \Pr(\mathbb{P}(\boldsymbol{X}_i, r, \lambda); \theta, M) \right) - \\
&\qquad nL \log(2)
\end{aligned} \tag{1.5}
$$

Our goal is to maximize (1.5) over θ, M, and $\mathcal{P}$ in order to estimate these parameters. One naive approach would be to generate all possible recent sample pedigrees and maximize the log-likelihood conditional on each pedigree in turn. However, the number of pedigrees to consider is prohibitively large even if only the first few generations back in time are considered. Many sample pedigrees will contain many IBD or admixture events and thus be unlikely to occur in nature, and in many populations, it is more probable that the sample have few IBD or admixture events in the very recent pedigree, if any. With this in mind, we consider only pedigrees containing no more than two events, whether they be IBD events or admixture events. Since we assume that we do not know the parental origin of each sequence, we further reduce the number of pedigrees to consider by evaluating only pedigrees that are unique up to labeling of ancestors as maternal and paternal.

In a two-deme population, each pedigree with two or fewer IBD or admixture events has the shape of one of the pedigrees shown in Figure S4. There are at most 21 distinct shapes of pedigrees with two or fewer events, and for each such pedigree shape, there exist some number of pedigrees with unique labelings of the sampled individuals and timings of the events in the pedigree. (Fewer distinct shapes are possible if the sample size is too small to permit certain shapes.) Table 1.1 gives the number of distinct pedigrees that must be considered for different sample sizes and numbers of past generations

Table 1.1: Number of distinct pedigrees with two or fewer IBD or admixture events. Pedigrees that differ only in maternal-paternal labeling of individuals are not counted as distinct.

generations	sample size			
	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$g = 2$	16	123	434	1109
$g = 3$	41	328	1144	2879
$g = 4$	78	631	2190	5477

considered. We note that we consider only pedigrees with non-overlapping generations and the sampled individuals found in the current generation. This is sufficient for the Wright-Fisher-like model of pedigrees that we investigate here, but real pedigrees will tend not to satisfy these criteria. Allowing overlapping generations will require consideration of many additional pedigrees and introduce problems of identifiability, with multiple pedigrees having the same set of reconfigurations with the same probabilities. We do not explore these issues here.

To calculate the standard sampling probabilities needed in (1.4), we use the method of Wu [31], which uses a dynamic-programming approach to efficiently calculate sampling probabilities of sequences generated under the infinite-sites mutation model in a two-deme population. In principle, it should be possible to calculate the log-likelihood of all pedigrees simultaneously, since any reconfiguration of the sample by recent IBD or admixture must correspond to one of the ancestral configurations in the recursion solved by Wu's [2010] method [see also 30]. Thus, for particular values of θ and M, after solving the ancestral recursion only once (and storing the sampling probabilities of all ancestral configurations), the likelihood of any pedigree can be found by extracting the relevant probabilities from the recursion. However, in order to take this approach to maximize the log-likelihood, it is necessary to solve the recursion on a large grid of θ

and M. In practice, we find that it is faster to maximize the log-likelihood separately for each pedigree, using standard derivative-free numerical optimization procedures to find the M and θ that maximize the log-likelihood for the pedigree. For simplicity, we use ordered sampling probabilities throughout, since the typical unordered probabilities would be inappropriate in this context due to the partial ordering of the sample by the pedigree.

To test our example inference method we simulated datasets of 1000 independently segregating loci generated by simulating coalescence through a randomly generated pedigree of a two-deme population with deme size $N = 1000$ and migration rate $M \in \{0.2, 2.0\}$. We sampled one individual (two sequences) from each deme. Sequence data were generated by placing mutations on the simulated gene genealogies according to the infinite-sites mutation model with rate $\theta = 1.0$ (when $M = 0.2$) or $\theta = 2.0$ ($M = 2.0$). In order to investigate the effects of admixture on the estimation on θ and M, each replicate dataset was conditioned upon having one of three different sample pedigrees with differing amounts of admixture (see Fig. 1.5). We calculated maximum-likelihood estimates of θ and M for each of the 41 distinct pedigrees containing two or fewer IBD or admixture events occurring in the past three generations. We compared these estimates to the estimates that would be obtained from a similar maximum-likelihood procedure that ignores the pedigree (i.e., assuming the null pedigree of no sample reconfiguration).

When there was recent admixture in the sample, assuming the null pedigree to be the true pedigree produced a bias towards overestimation of the migration rate (Fig. 1.5), since the early probability of migration via the admixture path must be accommodated by an increase in the migration rate. For this reason, the overestimation of the migration rate was greater when the degree of admixture was greater. The mutation rate was also

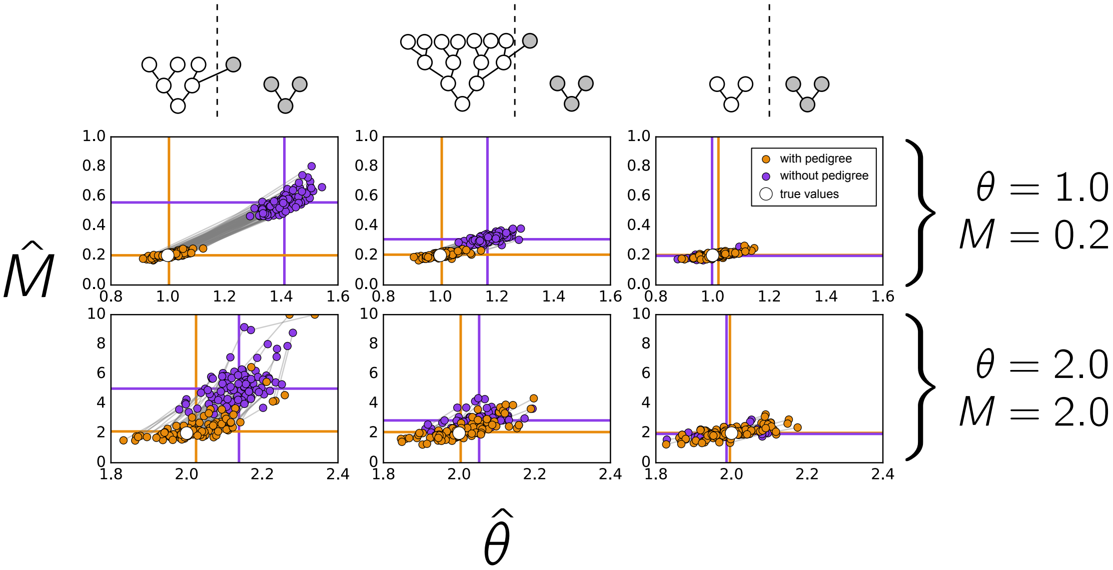

Figure 1.5: Maximum-likelihood mutation rates and migration rates for genetic datasets generated with different sample pedigrees. Each point depicts the maximum-likelihood estimates of θ and M for a particular simulation. Orange points show estimates obtained when the pedigree is included as a free parameter, and purple points show estimates obtained when the pedigree is assumed to have no effect on the data. In the first two columns, one of the sampled individuals is conditioned upon having a relative from the other deme, and in the third column the data are generated from completely random pedigrees. In each panel the true parameter values are shown with a solid white circle, and horizontal and vertical lines show means across replicates. Gray lines connect estimates calculated from the same dataset.

overestimated when the admixture in the pedigree was ignored, presumably because migration via the admixture path did not decrease allelic diversity as much as the overestimated migration rate should. Including the pedigree as a free parameter in the estimation corrected this biased estimation of M in the presence of admixed ancestry in the sample. Estimates from simulations of samples lacking any features in the recent pedigree produced approximately unbiased estimates of θ and M (Fig. 1.5).

The pedigree was not inferred as reliably as the mutation and migration rates (Fig. 1.6). When the simulated pedigree contained admixture, the estimated pedigree was the correct pedigree (out of 41 possible pedigrees) roughly half of the time. For pedigrees with no admixture and no IBD, the correct pedigree was inferred about one third of the time. In addition to calculating a maximum-likelihood pedigree, it is possible to construct an approximate 95% confidence set of pedigrees using the fact that the maximum of the log-likelihood is approximately χ^2 distributed when the number of loci is large. These pedigree confidence sets contained the true pedigree $\sim 88 - 99\%$ of the time, depending on the true sample pedigree, mutation rate, and migration rate. A log-likelihood ratio test has nearly perfect power to reject the null pedigree for the simulations with the lesser migration rate; for simulations with the greater migration rate the power depended on the degree of admixture, with more recent admixture producing greater power to reject the null pedigree (Fig. 1.6). Type I error rates for simulations where the null pedigree is the true pedigree were close to $\alpha = 0.05$.

We also simulated datasets of 1000 loci taken from samples with completely random pedigrees in a small two-deme population of size $N = 50$ per deme with one of two different migration rates ($M \in \{0.2, 2.0\}$). Whether or not the pedigree was included as a free parameter mostly had little effect on the estimates of θ and M (Fig. 1.7). However, in the few cases when the sample pedigree included recent admixture, the estimates were biased when the sampled pedigree was not considered as a free parameter. Inferring the pedigree together with the other parameters corrected this bias.

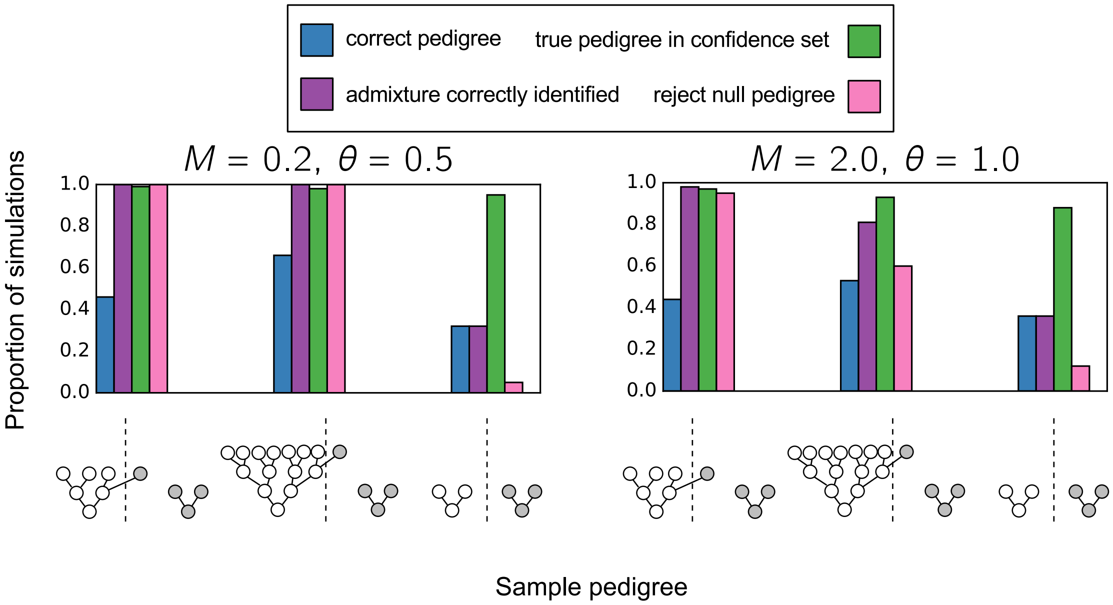

Figure 1.6: Inference of sample pedigrees. For simulations of 1000 infinite-sites loci, with $\theta = 0.5$ and $M = 0.2$ **(A)** or $\theta = 1.0$ and $M = 2.0$ **(B)**, different measurements of the accuracy and power of pedigree inference are shown. The conditioned-upon sample pedigrees are shown at the bottom of the figure. The blue bars show the proportion of simulations in which the maximum-likelihood pedigree was the true pedigree. The purple bars show the proportion of simulations where it was inferred that sampled individuals had admixed ancestry. The green bars show the proportion of simulations in which the true pedigree was found within the approximate 95% confidence set of pedigrees, and the pink bars show the proportion of simulations in which the null pedigree is rejected by a log-likelihood ratio test.

1.3 Discussion

Here we have explored the effects of migration events fixed in the population pedigree on the patterns of coalescence of unlinked loci. In contrast to the case of well-mixed, random-mating populations, in structured populations the population pedigree can influence coalescence well beyond the time scale of $\sim \log_2(N)$ generations in the past.

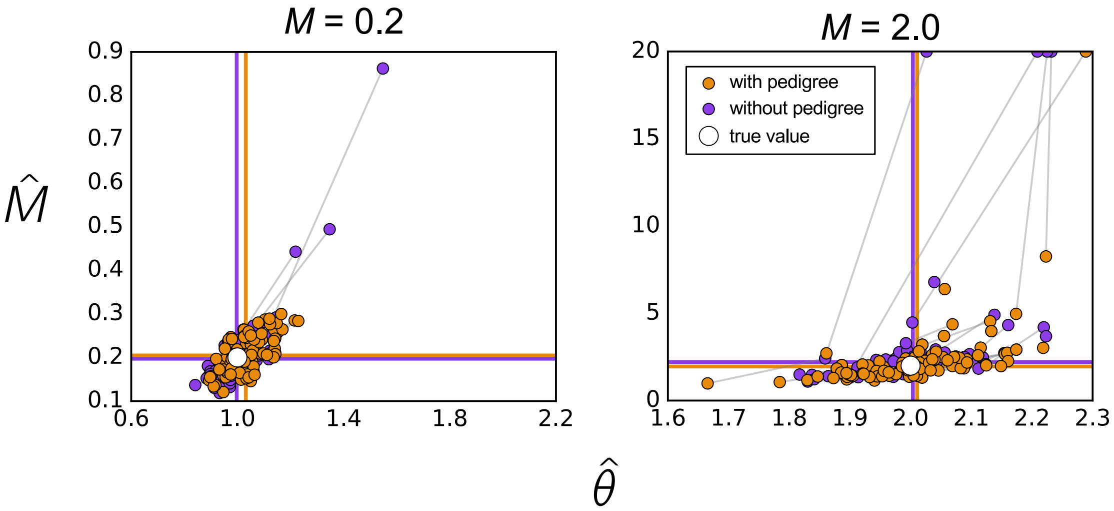

Figure 1.7: Maximum-likelihood mutation rates and migration rates for datasets of 1000 loci segregated through sample pedigrees simulated in a two-deme population model with deme size $N = 50$. Each point depicts the maximum-likelihood estimates of θ and M for a particular simulation. Orange points show estimates obtained when the pedigree is included as a free parameter, and purple points show estimates obtained when the pedigree is assumed to have no effect on the data. True parameter values are shown with white circles, and horizontal and vertical lines show means across replicates. Gray lines connect estimates calculated from the same dataset.

These effects are greatest when the total number of migration events occurring per generation is of the same order as the coalescence probability. When migration occurs more frequently than this, the particular history of migration events embedded in the population pedigree has less of an effect on coalescence, and the coalescence distributions based on the structured coalescent serve as good approximations to coalescent distributions within pedigrees.

We have also proposed an approach for incorporating the recent ancestry of the sample into coalescent-based inference of population mutation and migration rates in a two-deme population model. Unaccounted-for admixed ancestry of the sampled individuals introduces a bias in estimated migration rates, and this bias is eliminated by the inclusion of the sample pedigree as a free parameter in this inference.

The inference approach we have described above is in multiple ways complementary to existing procedures for inferring recent admixture and relatedness in structured populations. The popular program STRUCTURE [29] and related methods [32, 33, 34] are powerful and flexible tools for inferring admixture and population structure. Likewise, the inference tools RelateAdmix [35], REAP [36], and KING-robust [37] all offer solutions to the problem of inferring relatedness in the presence of population structure and admixture. Perhaps the most similar in scope is the method of Wilson and Rannala [38], which uses inferred ancestry proportions to estimate migration rates in the most recent generations. For input, each of these methods take genotypes at polymorphic sites, often biallelic SNPs, that are assumed to segregate independently. Likelihoods are calculated from the probabilities of observing the observed genotypes under the rules of Hardy-Weinberg equilibrium. These methods are well suited for samples of a large number of SNP loci sampled from a large number of individuals. The inference procedure we implemented, on the other hand, is capable of handling a sample of only a few individuals ($n \approx 4$), and we assume that the infinite-sites mutation model holds, with no recombination within loci and free recombination between loci. Likelihood calculations in our method are based on the coalescent in an explicit population genetic model, and the parameters of this model are the primary objects of inference. The pedigree is also explicitly modeled. We have shown that our approach works well when its assumptions hold. If the narrow assumptions of our model do not hold, or if the primary goal of inference is to infer recent features of the sample pedigree *per se*, other, more flexible methods are likely to be better.

Underlying our inference method is a hybrid approach to modeling the coalescent. Probabilities of coalescence are determined by the sample pedigree in the recent past, and then the standard coalescent is used to model the more distant past. This is similar to the approach used by Bhaskar et al. [39] to model coalescence when the sample size approaches the population size. In such a scenario, they suggest using a discrete-time Wright-Fisher model to model coalescence for the first few generations back in time and then use the standard coalescent model after the number of surviving ancestral lineages becomes much less than the population size. We note that in a situation where the sample size nears the population size in a diploid population, there will be numerous common ancestor events and admixture events in the recent sample pedigree, so it may be important to consider the pedigree when genetic variation is sampled from a fixed set of individuals at independently segregating loci.

Our sample reconfiguration framework could in theory be applied to models that allow the demography of the population to vary over time [e.g., 40, 41]. In such an application, if only a few individuals are sampled, it would be important to distinguish between the effects of very recent events that are particular to the sample and the effects of events that are shared by all individuals in the population. The latter category of events are more naturally considered demographic history. On the other hand, if a sizable fraction of the population is sampled, inferred pedigree features may be used to learn more directly about the demography of the population in the last few generations. As sample sizes increase from the tens of thousands into the hundreds of thousands and millions [1], it will become more and more possible to reconstruct large (but sparse) pedigrees that are directly informative about recent demographic processes.

Unexpected close relatedness is frequently found in large genomic datasets [e.g. 42, 43, 44]. It is common practice to remove closely related individuals (and in some cases, individuals with admixed ancestry) from the sample prior to analysis, but this unnecessarily reduces the amount of information that is available to make inferences. What is needed is a fully integrative method of making inferences from pedigrees and genetic variation, properly incorporating information about both the recent past contained in the sample pedigree and the more distant past that is the more typical domain of population genetic demographic inference. Here, by performing simulations of coalescence through pedigrees, we have justified a sample reconfiguration framework for modeling coalescence in pedigrees and given an example of how this can be incorporated into coalescent-based demographic inference. We hope this work spurs further investigations of pedigrees and patterns of coalescence.

2

The distributions of IBD segment lengths and the number of mutations separating IBD segments

2.1 Introduction

Identity by descent (IBD) is a central concept in the study of genetic variation. In the broadest sense, two genetic samples are IBD if they are identical due to coinheritance of genetic material. However, the definition of identity and the scope of coinheritance vary so much from usage to usage that IBD is possibly best considered to be an umbrella concept encompassing many loosely related concepts. As a concept, IBD has historically been used to describe patterns of coinheritance of Mendelian loci in pedigrees. It has also been used to describe allelic identity that is due to coinheritance rather than recurrent mutation. More recently, many authors have employed a concept of IBD defined in terms of non-recombinant chromosomal segments coinherited from a single ancestor. Thompson [45] provides a recent review of the various ways of thinking about IBD.

In this chapter we adopt the concept of IBD defined by recombination. We say that a segment of two aligned chromosomes is IBD if the segment spans contiguous ancestral material inherited from a single ancestor of the two chromosomes. Segments of identity by descent defined this way are delimited by recombination events that occurred during the ancestry of the sampled chromosomes. Under this definition of IBD, all chromosomes are IBD with all other chromosomes at all points along the chromosome; what varies with each pair of chromosomes is the distribution of breakpoints between IBD segments along the aligned chromosomes [see 45, 46, 47]. Most segments of IBD defined this way are difficult to detect in genetic data and are only reliably detected when the IBD segment is long (and thus relatively young). For this reason, previous studies employing this concept of IBD have considered only segments surpassing a given length to be IBD [reviewed by 48]. The use of a threshold length also provides a possibility of non-identity to complement the concept of identity, and it limits the considered timescale to the moderately recent past. However, genomic resources and IBD inference tools continue to improve, and it is becoming increasingly possible to detect shorter and shorter IBD segments [48].

Here, we investigate the full distribution of IBD segment lengths. To make theoretical predictions, we employ an assumption equivalent to the assumption used by Marjoram and Wall [11] to distinguish their sequentially Markov coalescent model (termed SMC') from the sequentially Markov coalescent model (SMC) of McVean and Cardin [10]. Both of these models provide an approximation to the coalescent with recombination (CwR), often referred to as the "ancestral recombination graph," which is the full, idealized model of ancestry for a sample of recombining chromosomes [30]. The SMC and SMC' simplify the CwR by disallowing certain coalescence events, which grants gives the pattern of ancestry across a chromosome the Markov property. In particular, the SMC allows coalescence events only between ancestral segments containing overlapping ancestral material. That is, two ancestral lineages are allowed to coalesce only if they carry genetic material at some site that is inherited by two individuals in the sample. Under the full CwR, all pairs of lineages are allowed to coalesce, regardless of whether ancestral material overlaps, so this is an approximation. However, the approximation is surprisingly good [10]. The SMC' modifies the SMC to allow coalescence between pairs of lineages containing overlapping *or adjacent* ancestral material. This improves the approximation to the CwR and retains the Markov property for patterns of ancestry across the chromosome [11].

Initial calculations of IBD segment lengths were based on the SMC [49]. Using the assumptions of the SMC', we calculate the marginal distribution of IBD segment lengths. We also calculate the distribution of IBD segment ages conditional on segment length, and we use this distribution to show that the age of many IBD segments typically found in the recent literature is too ancient to reflect any features particular to the pedigree of the sample. We conclude with an investigation the mutations on IBD segments and calculate the probability of true mutational identity for recombinational IBD segments of different lengths, finding that the number of mutations along a segment rapidly converges to a simple distribution (a Negative Binomial distribution) as the IBD segment length increases. Throughout, we find that predictions made with SMC' assumptions are indistinguishable from the results of simulations of the full CwR.

Some of the results in this chapter, in particular those related to lengths of IBD segments under the assumptions of the SMC', were independently derived by Shai Carmi, with whom I subsequently co-authored a paper [8]. Similarly, some of the results related to ages of IBD segments of different lengths and mutations on IBD segments were incorporated into a manuscript that I co-authored with Pier Palamara [9].

2.2 Results

Our approach to deriving distributions related to IBD segment lengths closely follows that of Palamara et al. [49]. Consider a chromosome of total genetic length r Morgans. Assume that in each generation, recombination occurs across the chromosome at rate r without interference. We sample two such chromosomes and choose a focal point along the aligned chromosomes. From the focal point we look to the right and to the left along the aligned chromosomes to determine the extent of recombinational IBD. Assume that the chromosome is long enough that its edges have little effect on the distribution of the length of the IBD segment containing the focal point. Suppose also that the coalescence time is t at the focal point, where the coalescence time has been rescaled by N, the haploid population size, in the coalescent limit. With time rescaled this way, ancestral lineages coalesce with rate 1.

At this point, recall that we define an IBD segment to be a chromosomal segment spanning contiguous ancestral material inherited from a single ancestor. This differs slightly from definitions of recombinational IBD used in previous studies, which define IBD in terms of the distance between nearest recombination events. In the context of the coalescent with recombination, two ancestral lineages formed by a single recombination event (i.e., two lineages ancestral to the same descendent chromosome) can coalesce back together, possibly prior to the coalescence event in the ancestor that defines the recombinational IBD segment. If such a “healing” or “back” coalescence event occurs, all evidence of the recombination event is essentially erased, and the ancestral material comprising the IBD block extends to a subsequent recombination point. This line of argument is equivalent to the modification of the sequentially Markov coalescent (SMC) into the SMC’, in which the new lineage formed by recombination is allowed to coalesce back onto the gene genealogy that is adjacent along the chromosome [11].

When time is measured in discrete generations, recombination events occur across the chromosome at a rate of 1 per Morgan in each generation. When time is modeled as continuous and rescaled by N, recombination happens at a rate of $\rho/2dt = Ndt$ per Morgan in the infinitesimal time interval $(t, t + dt)$. (In this chapter we will measure segment lengths in units of ρ; to convert to Morgans, divide the segment length by $2N$.) Looking to the right from the focal point on a single chromosome, if the focal coalescence time is t, the total exponential rate of arrival for the nearest recombination event is $\int_0^t \rho/2du = \rho t/2$. When considering the nearest recombination event along a pair of chromosomes, this rate is ρt. However, at each time s, $0 < s < t$, recombination events at that time produce two ancestral lineages that coalesce back together prior to t with probability $\frac{1}{2}(1 - e^{-2(t-s)})$. To see that this is the case, consider the three possible outcomes at time t after a recombination event occurs at time $s < t$ (Fig. 2.1). The three possible outcomes are (1) no coalescence occurs prior to time t; (2) the broken-off lineage (i.e., the lineage not containing the focal point) coalesces back onto the lineage containing the adjacent ancestral material, and (3) the broken-off lineage coalesces with ancestral lineage containing the homologous positions sampled from the other chromosome. The probability of no coalescence is $e^{-2(t-s)}$, since each of the two coalescence events occurs at rate 1. Given that a coalescence event does occur, the two

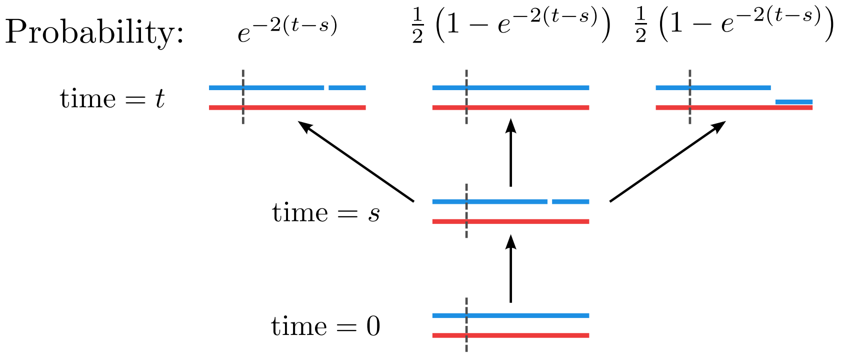
Probability:
$e^{-2(t-s)}$
$\frac{1}{2}\left(1-e^{-2(t-s)}\right)$
$\frac{1}{2}\left(1-e^{-2(t-s)}\right)$
time $= t$
time $= s$
time $= 0$

Figure 2.1: The three possible outcomes of a recombination event that occurs at time s prior to a conditioned-upon local coalescence time t. After the recombination event occurs, the broken-off lineage can coalesce either back onto the adjacent ancestral material or onto the homologous ancestral material. Each of these coalescence events occurs at rate 1, so the probability that no coalescence event occurs is $e^{-2(t-s)}$. The probability that each coalescence event occurs is $\frac{1}{2}(1-e^{-2(t-s)})$.

coalescence events are equally probable, so that the probability that the recombination event back-coalesces before time t is $\frac{1}{2}(1-e^{-2(t-s)})$.

Since we are interested only in the distance to the nearest recombination event causing a change in most recent common ancestry along two chromosomes, in the time interval $(s, s+ds)$, $0 < s < t$, the *effective* recombination rate across the chromosome is $\rho[1-\frac{1}{2}(1-e^{-2(t-s)})]ds$. Thus the total rate of arrival of the first effective recombination event (*i.e.*, the first change in the most recent shared ancestor) between times 0 and t is

$$\int_0^t \rho\left[1-\frac{1}{2}\left(1-e^{-2(t-s)}\right)\right]ds = \frac{\rho}{4}\left(1-e^{-2t}+2t\right).$$

Looking to the left from the focal point, if we make the same assumption that is used to derive the SMC', namely that ancestral lineages containing exclusively non-overlapping and non-adjacent ancestral material cannot coalesce [11], then the process of generating the extent of recombinational IBD on the left of the focal point is independent of the process of generating the extent of recombinational IBD on the right. With this assumption, the distribution of the length L of the IBD segment containing the focal point is Erlang-2 $\left(\frac{1}{4}\left(1-e^{-2t}+2t\right)\right)$, with density

$$f_{L|T}(l|t) = \frac{l}{16} e^{-\frac{1}{4}l\left(1-e^{-2t}+2t\right)} \left(1-e^{-2t}+2t\right)^2. \tag{2.1}$$

If the IBD segment is defined by adjacent sites of ancestral recombination, instead of contiguous ancestral material, the total rate of arrival of the IBD segment boundary is ρt and the distribution of L^* is Erlang-2(t). This is equivalent to assuming the SMC model of coalescence and recombination, since under the SMC, each recombination event produces a new ancestor between a pair of chromosomes. (We will use asterisks to mark random variables related to IBD under the SMC assumptions.) For random variables X and Y, X is defined to be stochastically less than Y if $\Pr(X > a) \leq \Pr(Y > a)$ for all a. Because the distribution of IBD segment lengths under the SMC does not account for back coalescence events, it is stochastically less than the length distribution under the SMC'.

Figure 2.2 shows the difference between the conditional length distributions for L and L^*. When the focal coalescence time is short, there is little difference between the distributions because there is little time for additional coalescence events after the recombination event. When the focal coalescence time is longer, there is more opportunity for a back coalescence event and the distributions are less similar.

To derive the marginal distribution of IBD segment lengths unconditional on local pairwise coalescence time, we integrate $f(l|t)$ over t:

$$f_L(l) = \int_0^\infty f_{L|T}(l|t) f_T(t) dt = \int_0^\infty \frac{l}{16} e^{-\frac{1}{4}l\left(1-e^{-2t}+2t\right)} \left(1-e^{-2t}+2t\right)^2 e^{-t} dt. \tag{2.2}$$

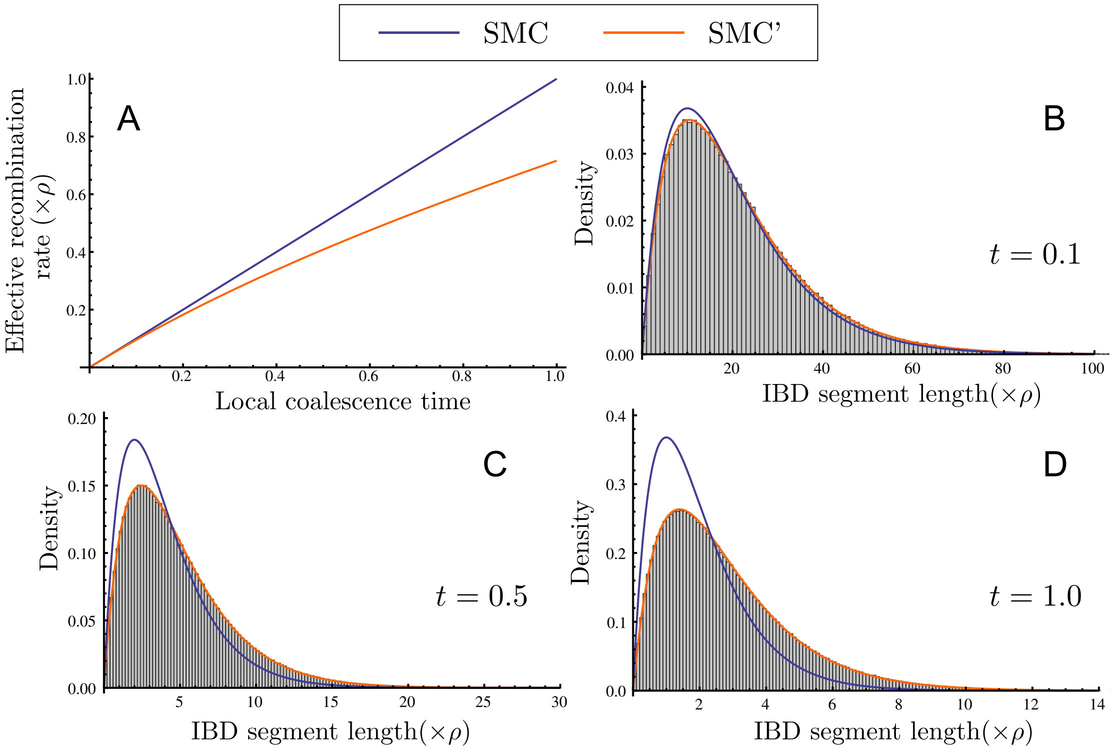

Figure 2.2: Comparison of the distribution of IBD lengths given a local coalescence time. In panel **A**, the total effective recombination rate is shown as a function of local coalescence time. Under the SMC, the total recombination rate grows linearly with coalescence time. Under the SMC', the total effective recombination rate is less for older segments because they have a greater tendency to produce recombination events that have a substantial probability of back coalescence. In panels **B**–**D**, comparisons of Erlang-2 length distributions for different local coalescence times are shown. When the local coalescence time is greater, the effect of back coalescence is greater and the SMC and SMC' distributions are more different. Histograms show results of simulations of the full coalescent with recombination carried out in `ms`. In each panel, the simulation results are the IBD segments whose ages are in the interval $[0.95t, 1.05t]$ taken from a simulation of 50 million IBD segments, each at the midpoint of a chromosome of total length $\rho = 1000$.

The solution of this integral can be expressed in terms of special functions and is given in Appendix D. The solution of the equivalent integral for L^* is

$$f_{L^*}(l) = \int_0^\infty f_{L^*|T}(l|t) f_T(t) dt = \frac{2l}{(1+l)^3}.$$

This distribution has no moments, and because L is stochastically greater than L^*, the distribution of L must also have no moments. Figure 2.3A shows the difference between these two distributions and compares each to the distribution of IBD segments under the coalescent with recombination.

This approach to deriving the distributions of L and L^* corresponds to a particular process of sampling IBD segments. Since derivations are in reference to a given focal point, the sampling process is equivalent to choosing a point (rather than a segment) uniformly at random from a collection of IBD segments arranged sequentially across an infinite chromosome. By choosing a point at random, segments are sampled with weight proportional to their lengths; this is known as the inspection paradox. If instead the segments are weighted equally and a segment is sampled uniformly at random, the density of such a segment S is given by the following integral:

$$f_S(s) = \frac{\frac{1}{s} f_L(s)}{\int_0^\infty \frac{1}{u} f_L(u) du} = \frac{3 f_L(s)}{2s}. \tag{2.3}$$

Note that the kth moment of S is $3/2$ multiplied by the $(k-1)$th root of L. Thus $\mathrm{E}[S] = 3/2$ and S has no higher moments. The equivalent density for S^* is

$$f_{S^*}(s) = \frac{2}{(1+s)^3}.$$

This distribution also has a mean ($\mathrm{E}[S^*] = 1$) but no higher moments. Figure 2.3B

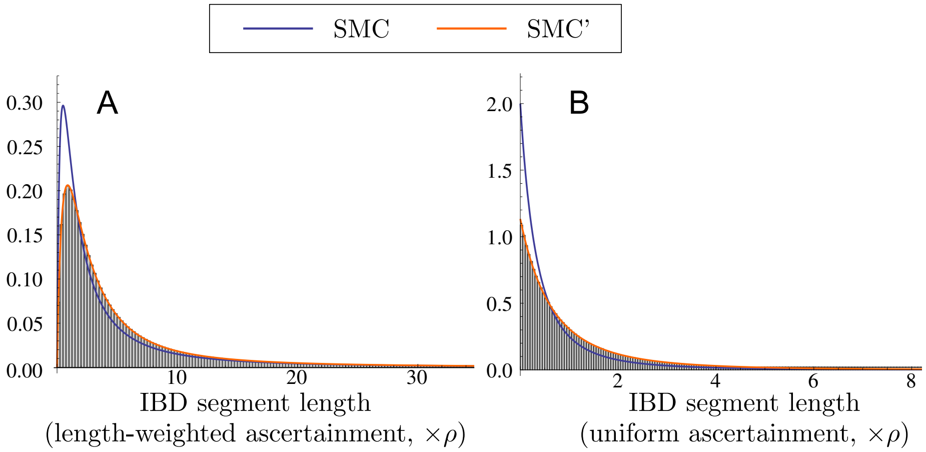

Figure 2.3: Comparison of the distribution of IBD lengths under the SMC, SMC', and CwR. Panel **A** shows the distribution of IBD segment lengths when segments are sampled with weight proportional to their length (i.e., IBD segments encompassing a fixed point). Panel **B** shows the distribution of IBD segment lengths when segments are sampled uniformly at random. In both ascertainment schemes, IBD segments are longer when back coalescence is modeled, *i.e.,* under the SMC' assumptions. Histograms show results of simulations of the full coalescent with recombination carried out in `ms`. Simulation sample sizes are 5 million for both panels.

compares the distributions of S and S^*.

Using Bayes' Rule, it is possible to obtain the distribution of pairwise coalesce times given an IBD segment length. This is

$$f_{T|L}(t|l) = \frac{e^{-t} f_{L|T}(l|t)}{f_L(l)}. \tag{2.4}$$

Figure 2.4 shows this distribution for different IBD segment lengths. It is notable that for segments of lengths typically inferred from recent studies, for example segments of length in the range of 1 to 5 cM, the distribution of ages of these segments extends greatly beyond the $\log_2(N)$ timescale on which the population pedigree converges [7, 26, see $t = \log_2(N)/N \approx 0.0013$ for $N = 10000$ in Figure 2.4]. This shows that most IBD segments in this length range are due to coalescence events in ancestors that are shared equally by all individuals in the population; thus the presence of such recombinational

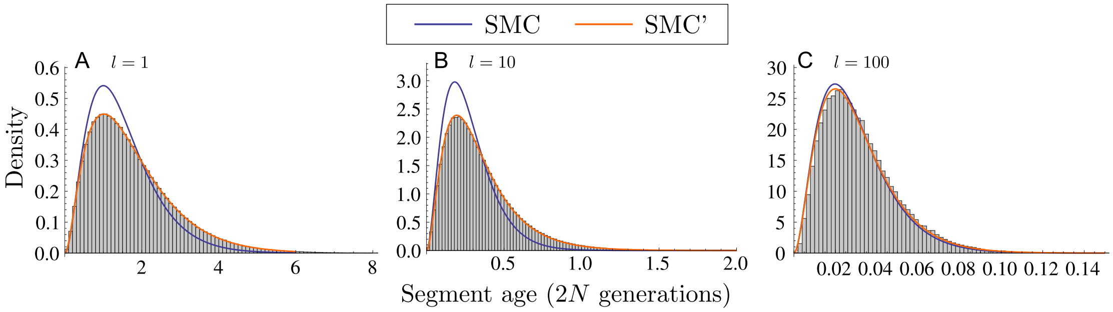

Figure 2.4: Distribution of IBD segment ages conditional on segment length. Panels **A**–**C** show ages of IBD segments of lengths $l = 1$, $l = 10$, and $l = 100$, respectively. In each panel, a histogram shows simulated ages of IBD segments in the length interval $[0.95l, 1.05l]$. The simulation results taken in each panel were taken from a large pool ($n = 5$ million) of random IBD segment lengths simulated using `ms` with a total recombination rate of $\rho = 1000$. Sample sizes were $n = 1023920$ (**A**), $n = 929020$ (**B**), and $n = 99806$ (**C**). There were many fewer simulations for panel **C** ($l = 100$) because IBD segments of that length are relatively rare. Theoretical predictions in panel **C** are less accurate presumably because of the effects of finite chromosome length ($\rho = 1000$) on sampling a segment of length $l = 100$.

IBD segments conveys little information about the pedigree relationships particular to the two individuals possessing the pair of chromosomes being compared, if such particular relationships exist.

Mutation

The calculations above can be used to study mutational differences along recombinational IBD segments. Assume that mutation occurs according to an infinite-sites model at rate $\theta/2$. Because we have scaled lengths by ρ, the relevant parameterization of mutation is the ratio $\nu = \theta/\rho$. Conditional on the length l and coalescence time t of an IBD segment, the number of heterozygous sites along the segment is Poisson distributed with mean $\nu l t$. Thus the joint distribution of the local coalescence time, IBD segment length, and the number of heterozygous sites on the IBD segment is

$$f_{T,L,K}(t,l,k) = e^{-t} f_{L|T}(l|t) \frac{(\nu l t)^k}{k!} e^{-\nu l t}. \tag{2.5}$$

One can also derive the distribution of the number of mutations along an IBD segment given its length. This has the following form:

$$\Pr(K = k | L = l) = \int_0^\infty f_{T|L}(t|l) \frac{(\nu l t)^k}{k!} e^{-\nu l t} dt. \tag{2.6}$$

The integral in (2.6) can be solved numerically.

It is worthwhile to make the same calculations with the simpler expressions involving L^*, since L converges in distribution to L^* as $L \to \infty$. Under the SMC, the equivalent of equation (2.6) has the following solution:

$$\Pr(K = k | L^* = l) = \frac{1}{2}(k+1)(k+2) \left(\frac{1+l}{1+l+l\nu} \right)^3 \left(\frac{l\nu}{1+l+l\nu} \right)^k. \tag{2.7}$$

As $l \to \infty$, this becomes

$$\frac{1}{2}(k+1)(k+2) \left(\frac{1}{1+\nu} \right)^3 \left(\frac{\nu}{1+\nu} \right)^k, \tag{2.8}$$

which is a Negative Binomial distribution with success probability $\nu/(1+\nu) = \theta/(\theta+\rho)$ and $r = 3$. It is clear that equation (2.7) is well approximated by the limiting distribution (2.8) when $l \gg 1$. Equivalently, if the minimum considered IBD segment length is x cM, then (2.8) should provide a reasonably accurate approximation to the distribution of the number of mutations along IBD segments so long as $N \gg \frac{100}{x}$. This convergence argument is made with reference to equations involving L^* rather than L, but convergence of the number of mutations on L to equation (2.8) is observed on the same scale through numerical calculation of equation (2.6) (Fig. 2.5).

The concept of recombinational IBD is often linked to mutational identity through the assumption that large IBD segments inherited recently from a common ancestor will tend to be allelically identical. This assumption can be examined using the above calculations. When a segment is long, the Negative Binomial distribution given in equation (2.8) can be used, and the probability of complete mutational identity is $(1+\nu)^{-3} = \rho^3/(\theta+\rho)^3$. When the mutation and recombination rates are equal, this probability is 1/8. To consider the probability of complete identity for shorter segments, we can solve the integral in equation (2.6) with $k = 0$. This integral can be solved analytically for segments modeled under both the SMC and SMC', but the solution for segments modeled by the SMC' is long and involves special functions, so it is omitted. Under the SMC, the probability of complete identity is

$$\Pr(K = 0 | L^* = l) = \frac{(1+l)^3}{(1+l+l\nu)^3}.$$

This function decreases monotonically from unity to $(1+\nu)^{-3}$ as l increases from zero; thus, shorter segments always have a greater probability of being completely identical than longer segments. Interestingly, the probability of complete identity for IBD segments under the SMC' is minimized between zero and infinity. There is no simple expression for this minimum, but numerical calculations indicate that it grows sublinearly with ν. Figure 2.6 shows the probability of complete mutational identity as a function of IBD segment length, modeled by both the SMC and SMC'.

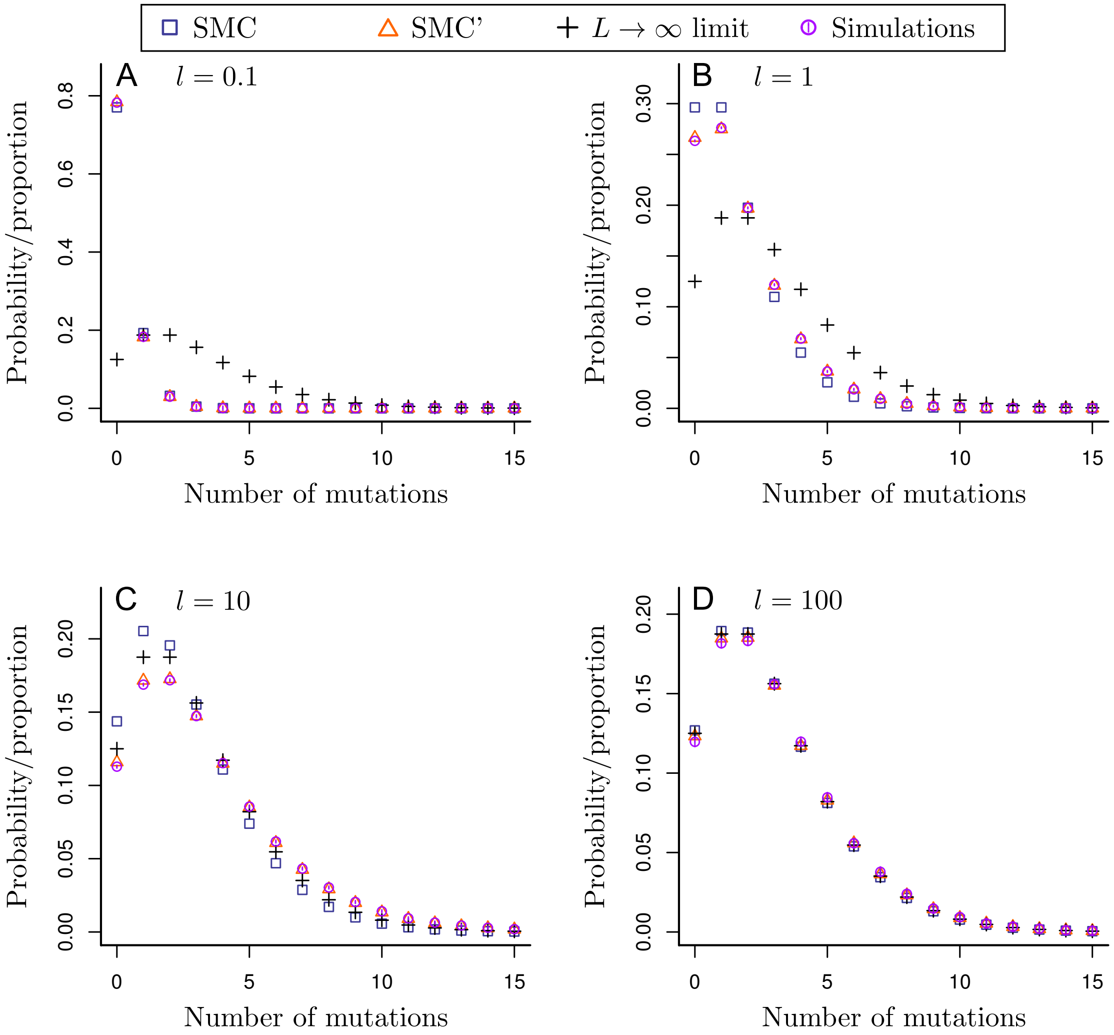

Figure 2.5: Distribution of the number of mutational differences along IBD segments of different lengths. Panels **A–D** show the distribution of the number of singleton mutations along IBD segments of lengths $l = 0.1, 1, 10$, and 100. In a diploid population of size 1000, for example, these lengths would correspond to 0.0025 cM, 0.025 cM, 0.25 cM, and 2.5 cM, respectively. Each panel shows theoretical predictions based on the length of the segment (modeled under the SMC and SMC'), predictions for the case where $L \to \infty$, and observations from simulations. Simulations were carried out in `ms`. Each simulation ($n = 50$ million) generated a chromosome with end-to-end recombination and mutation rates $\rho = \theta = 1000$. Vertical lines in the magenta circles show the standard errors of estimates from simulations.

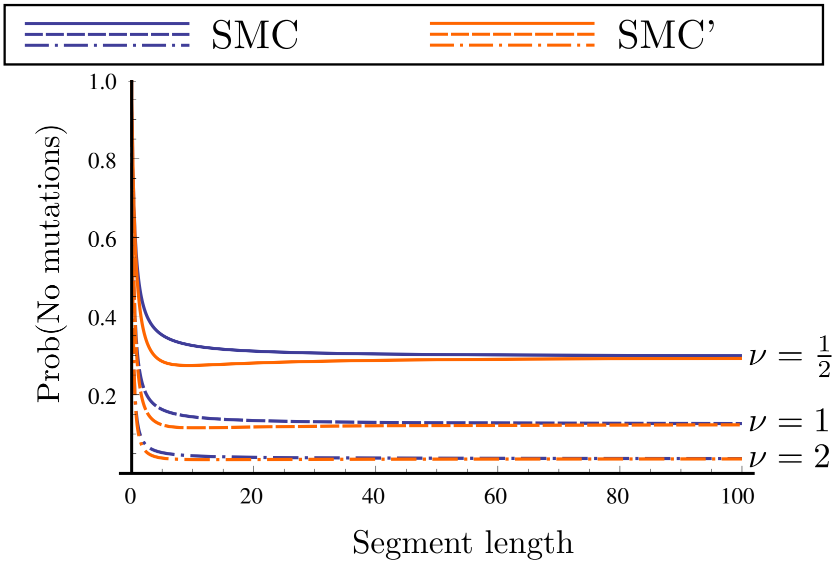

Figure 2.6: Probability of complete mutational identity as a function of segment length. Probabilities are shown for three different values of $\nu = \theta/\rho$, modeled by the SMC and SMC'. In each case the asymptotic probability of complete mutational identity is $(1+\nu)^{-3} = \rho^3/(\theta+\rho)^3$. Segment lengths are measured in units of $\rho = 2Nr$.

Simulations

To check calculations and determine the parameter space in which it is important to model back-coalescence events, we ran simulations using the coalescent simulator `ms`. When `ms` is provided with the `-T` option, it outputs local coalescence times that can be used to extract recombinational IBD segments that are generated by the full coalescent with recombination. In analyzing the output from `ms`, two segments broken by recombination were considered to be part of the same recombinational IBD segment if they were contiguous and had the same coalescence time. Because `ms` simulates the full coalescent with recombination, the size of the simulated chromosomes is limited by computational resources to a maximum total recombination rate of $\rho \approx 1000$, which is 5 cM in a haploid population of size 10000. This finite chromosome length likely introduces the effects of chromosome ends on longer simulated segments, and this is apparent in the data (for example, see Fig. 2.4C).

2.3 Discussion

Here, we have calculated various distributions related to recombinational IBD segments. We showed that the SMC' provides accurate predictions for the lengths of IBD segments across the whole range of segment lengths and ages. We also calculated the distribution of the age of an IBD segment given its age and and the number of mutations on an IBD segment given its length. As noted above, these results were incorporated into studies that I co-authored [8, 9].

IBD provides a powerful lens for investigating the recent history of a population. Palamara et al. [49] developed a method for reconstructing population sizes in the recent past based on observed IBD segment lengths and estimated recent bottlenecks and contractions in the Ashkenazi Jewish and Kenyan Maasai populations. Their method uses length distributions similar to those presented above, except they allow variable population sizes and implicitly base their calculations on the SMC rather than the SMC'. Ralph and Coop [47] studied IBD across Europe taking a more empirical approach and similarly based calculations on the SMC. Calculations based on the SMC are quite accurate for IBD segments with length greater than ~ 100 measured in units of $\rho = 2Nr$; for $N = 5000$, this is 0.5 cM. At the present, most studies consider only IBD segments longer than $\sim 1-2$ cM, since IBD detection methods struggle to accurately identify IBD segments shorter than this [but see 50, 51]. If the population is small or if shorter IBD segments are retained, it will be important to base calculations on the SMC' rather than the SMC. Short IBD segments are difficult to delimit, due to a lack of polymorphic sites or the possibility of multiple contiguous short segments being combined into a single segment. This presents an additional challenge beyond choice of SMC versus SMC'. In order to avoid these difficulties, Harris and Nielsen [52] used the lengths of intervals between sequence mismatches to infer demographic history in humans. They found that it was necessary to base calculations on the SMC' rather than the SMC in order to make accurate inferences.

3

The SMC' is a highly accurate approximation to the ancestral recombination graph

3.1 Introduction

Of the many models of genetic variation in the field of theoretical population genetics, few have as much relevance in the era of genomics as the ancestral recombination graph (ARG). The ancestral recombination graph models patterns of ancestry and genetic variation within sequences experiencing recombination under neutral conditions [53, 54].

Under the formulation of Griffiths and Marjoram [54], lineages recombine apart and coalesce together back in time to produce a graph structure describing the ancestral genealogy at each point along a continuous chromosome. While only a few simple rules govern the process, many aspects of the model are analytically intractable.

Wiuf and Hein [55] provided a formulation of the ARG that proceeds across the chromosome (rather than back in time), producing the genealogy at each point sequentially. As with the back-in-time formulation of Griffiths and Marjoram [54], at each point along the chromosome there is a local genealogy describing the ancestry of the sample at that point, and changes in the genealogy occur at points where recombination events have occurred. In this sequential formulation of the ARG, a new lineage is produced wherever an ancestral recombination event is encountered along the chromosome. To produce a new genealogy at the recombination site, the new lineage is coalesced to the ARG representing the ancestry of all previous points along the chromosome. This dependence on all previous points makes the process non-Markovian and is one of the properties of the ARG that makes it often intractable.

Approximations to the ARG have been suggested with the goal of modeling coalescence with recombination in a way that is analytically tractable. McVean and Cardin [10] introduced the *sequentially Markov coalescent* (SMC). The original formulation of the SMC was a sequential model, generating genealogies along the chromosome such that each new genealogy depends only on the previous genealogy. Like the ARG, the SMC has both a back-in-time formulation and a sequential formulation. The back-in-time formulation of the SMC is equivalent to that of the ARG except that coalescence is allowed only between lineages containing overlapping ancestral material. As a consequence, in the sequential formulation of the pairwise ($n = 2$ chromosomes) SMC, each recombination event produces a new pairwise coalescence time.

Marjoram and Wall [11] introduced a slight modification to the SMC, termed the SMC', which retains the Markov behavior along the chromosome but models additional coalescence events that make it a closer approximation to the ARG. Specifically, in the back-in-time formulation of the SMC', coalescence is allowed between lineages containing either overlapping *or adjacent* ancestral material. In the sequential formulation of the pairwise SMC', this means that not every recombination event necessarily produces a change in local coalescence time, since two lineages created by a recombination event can coalesce back together. Figure 3.1 shows the transitions that are permitted under the back-in-time and sequential formulations of the pairwise ARG, SMC, and SMC'. The sequentially Markov coalescent models have been used in many recently introduced population-genetic, model-based inference procedures, including the pairwise SMC (PSMC) model [12], multiple SMC (MSMC) model [17], diCal [13], coalHMM [16, 56], and ARGWeaver [14].

The SMC' was shown by simulation to produce measurements of linkage disequilibrium more similar to the ARG than those produced by the SMC [8, 11]. Few other comparisons between these models have been made, and analytical results for the SMC' are few. Here, we propose a model for generating pairwise coalescence times at two fixed points along continuous chromosomes modeled by the SMC'. Through analysis of this model, we calculate for the first time many statistical properties of the pairwise SMC' and compare these against those of the ARG and SMC. Specifically, for each model of coalescence with recombination, we compare the following properties of (T_1, T_2), the joint distribution of pairwise coalescence times at two fixed points: the joint density $f_{T_1,T_2}(t_1, t_2)$ (Section 3.2.2), the conditional density $f_{T_2|T_1}(t_2|t_1)$ (Section 3.2.3), the probability $P(T_1 = T_2)$ that the pairwise coalescence times are the same (Section 3.2.4), and the covariance $\text{Cov}[T_1, T_2]$ between the coalescence times (Section 3.2.5). These

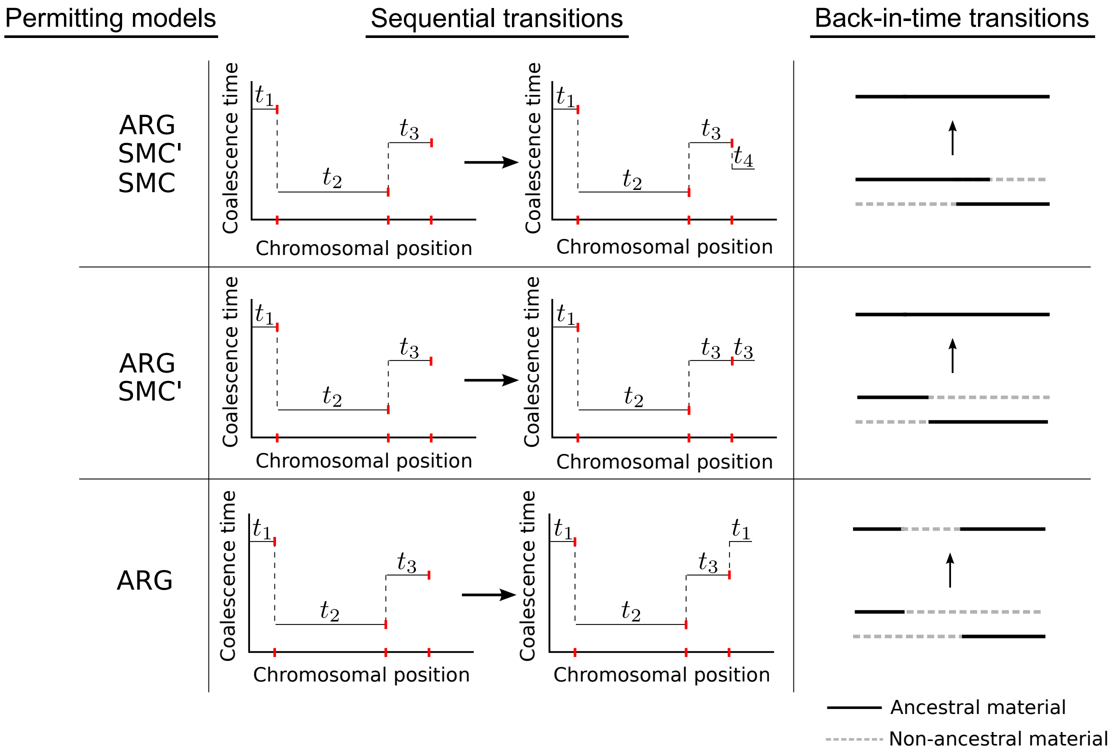

Figure 3.1: Transitions permitted under the ARG, SMC', and SMC models. Under "Sequential Transitions," a transition occurs from left to right across the chromosome at the rightmost recombination event. Locations of recombination events are marked with red lines along the horizontal axis, and the ith coalescence time is labeled as t_i. Under "Back-in-time transitions," the arrow indicates a coalescence event that occurs between two aligned ancestral chromosomes, each carrying a combination of ancestral (solid black line) and non-ancestral material (dashed gray line). Ancestral material is defined as a portion of a chromosome that is ancestral to the sample.

quantities are readily related to measures of linkage disequilibrium in real sequence data.

Using our two-locus model of the pairwise SMC', we also show that the joint distribution of coalescence times immediately to the left and right of a recombination event is the same under the SMC' and ARG. This allows us to calculate the asymptotic bias of the pairwise SMC- and SMC'-based population-size estimators, which we confirm by simulation. We show that the SMC' estimator is approximately asymptotically unbiased.

3.2 Results

3.2.1 Two-locus Markov chain models

We first present two-locus, back-in-time formulations (as opposed to sequential, across-the-chromosome formulations) of the ARG, SMC, and SMC' models, beginning with the previously described ARG and SMC because these models are simpler and provide context for the presentation of the SMC'. Kaplan and Hudson [57] described how the process of generating coalescence times at two linked loci modeled by the ARG can be represented as a continuous-time Markov chain, with coalescence and recombination events causing transitions between states. Simonsen and Churchill [58] explored this process further for the case where the sample size is $n = 2$ and derived for the ARG many of the quantities we compare against the SMC' in this paper.

If time is scaled such that the rate of coalescence is one and the total rate of recombination between the two linked loci is $\rho/2$, then the two-locus ancestral process under the ARG is the model depicted in Fig. 3.2A. The process starts in state $\boldsymbol{R_0}$ with two lineages, each containing linked copies of the two loci. From $\boldsymbol{R_0}$, the process transitions with rate ρ to state $\boldsymbol{R_1}$, in which one of the two chromosomes has experienced a recombination event, or to state $\boldsymbol{C_B}$, in which both loci have coalesced, terminating the process. From $\boldsymbol{R_1}$, a recombination event on the remaining linked chromosome (occurring with rate $\rho/2$) can take the process to $\boldsymbol{R_2}$, in which neither locus has coalesced and all focal-locus copies are unlinked, or to $\boldsymbol{C_L}$ or $\boldsymbol{C_R}$, in which the left and right focal loci have coalesced, respectively.

Unlike previous descriptions of two-locus continuous-time Markov chains [58, 59], we disregard any information about linkage between the two loci after one locus has coalesced, since the rate of coalescence at the uncoalesced locus is 1 regardless of the state of linkage with the coalesced locus. In general, the rate of coalescence between each pair of lineages containing uncoalesced ancestral material is 1, and the rate of recombination, breaking apart linked loci back in time, is $\rho/2$ multiplied by the number of lineages containing linked loci.

The defining feature of the SMC is that ancestral lineages not containing overlapping ancestral material cannot coalesce [10]. For the back-in-time formulation of the SMC, the consequence of restricting coalescence in this way is that once a recombination event occurs between the two loci, the process can never return to the fully-linked, uncoalesced state ($\boldsymbol{R_0}$) and the remaining time until coalescence at the two loci can be modeled as independent exponential random variables with rate 1 (Fig. 3.2B). This suggests a natural representation of the joint distribution of (T_1, T_2) under the SMC:

$$(T_1, T_2) \sim (X_0 + RX_L,\ X_0 + RX_R), \tag{3.1}$$

where $X_0 \sim \text{Exp}(1+\rho)$ is the amount of time to leave $\boldsymbol{R_0}$, $R \sim \text{Bernoulli}(\frac{\rho}{1+\rho})$ indicates whether the first event is a recombination event, and $X_L \sim X_R \sim \text{Exp}(1)$ are the exponential waiting times until coalescence after the first recombination event. All of these random variables are independent in the SMC model, so it is straightforward to calculate many of the quantities we compare in this paper using this representation.

The defining rule of the SMC' model of coalescence with recombination is that only ancestral lineages containing overlapping *or contiguous* ancestral material can coalesce [11]. The back-in-time model of coalescence at two fixed loci under this model is the Markov jump chain shown in Figure 3.3. Under the SMC', it is necessary to model the number of recombination events that have occurred between the two loci at each point in time. To see that this is the case, consider the state $\boldsymbol{R_2}$ in Figure 3.3. In this state, two recombination events have occurred between the focal loci, and neither focal locus has coalesced. Because lineages can only coalesce to lineages containing overlapping or adjacent ancestral material, two particular coalescence events would need to occur before the process returns to state $\boldsymbol{R_0}$, regardless of the placement of the recombination events on the two chromosomes. This model also features an additional state $\boldsymbol{I}$, which is entered when some portion of the chromosome between the focal loci coalesces prior to either of the focal loci. Upon entering $\boldsymbol{I}$ it becomes impossible for the process to re-enter the initial, fully-linked state ($\boldsymbol{R_0}$), so the remaining times until coalescence at the focal loci become independent exponential random variables with mean 1. If $\boldsymbol{R_i}$ is the state in which neither focal locus has coalesced and i recombination events have occurred between the focal loci, the transition rate into $\boldsymbol{I}$ is $i - 1$. This is due to the fact that each recombination event after the first produces an additional pair of lineages that can coalesce to take the process to $\boldsymbol{I}$. For each state $\boldsymbol{R_i}$, $i \geq 1$, the number of lineages that can coalesce to take the process to $\boldsymbol{R_{i-1}}$ is i, and the rate of transitioning to $\boldsymbol{R_{i+1}}$ through recombination is ρ. As with the ARG and SMC, transitions to $\boldsymbol{C_L}$ and $\boldsymbol{C_R}$ occur at rate 1 whenever the process is in state $\boldsymbol{R_i}$, $i \geq 1$.

3.2.2 Joint probability density functions

For the ARG, SMC, and SMC', let $R_0(t)$ represent the probability that the two-locus ancestral coalescent process is in state $\boldsymbol{R_0}$ at time t, and let $R^+(t)$ represent the probability that the process is in any state $\boldsymbol{R_i}$ at time t, where $i \geq 1$ (including $\boldsymbol{I}$ for the SMC'). For the three coalescent models we compare here, the general form of the joint density of coalescence times at the two focal loci is

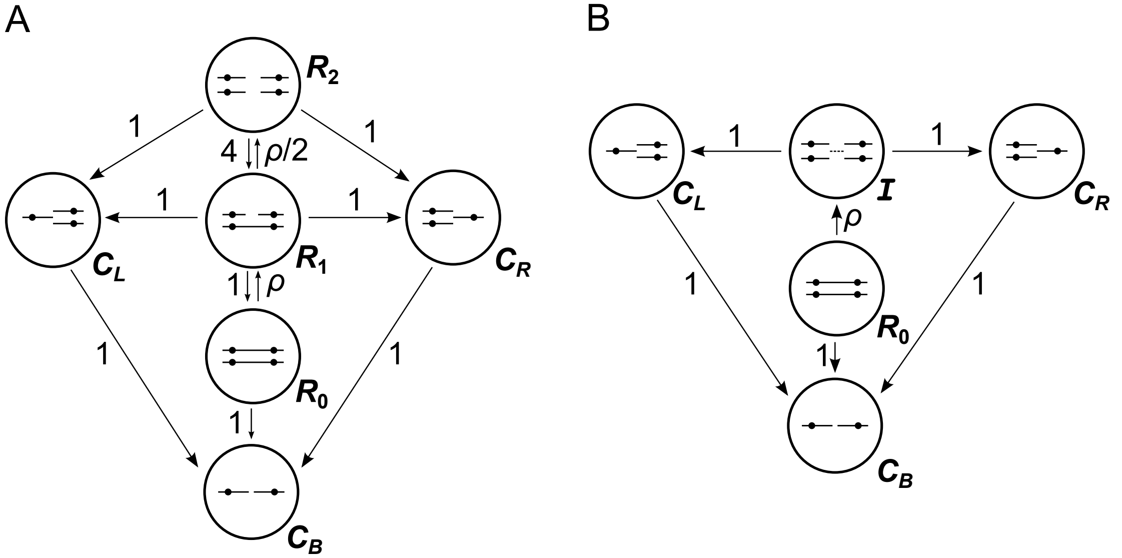

Figure 3.2: Panel **A** shows a schematic of the ARG back-in-time Markov process for two loci. Panel **B** shows schematic of the SMC back-in-time Markov jump chain for two loci. In both cases the process starts in state $\boldsymbol{R_0}$ and transitions to other states occur with the rates indicated by arrows between states.

$$f_{T_1,T_2}(t_1,t_2) = \begin{cases} R_0(t_1) & t_1 = t_2 \\ R^+(t_1)e^{-(t_2-t_1)} & t_1 < t_2 \\ R^+(t_2)e^{-(t_1-t_2)} & t_1 > t_2. \end{cases} \tag{3.2}$$

For the ARG and the SMC, the number of states is finite and $R_0(t)$ and $R^+(t)$ can be solved using matrix exponentiation. For the SMC', there are an infinite number of states, representing the possibility of an infinite number of recombination events occurring between the two focal loci. To solve for the probability $R_j(t)$ that the SMC' process is in state $\boldsymbol{R_j}$ at time t, one can use the forward Kolmogorov equation (for

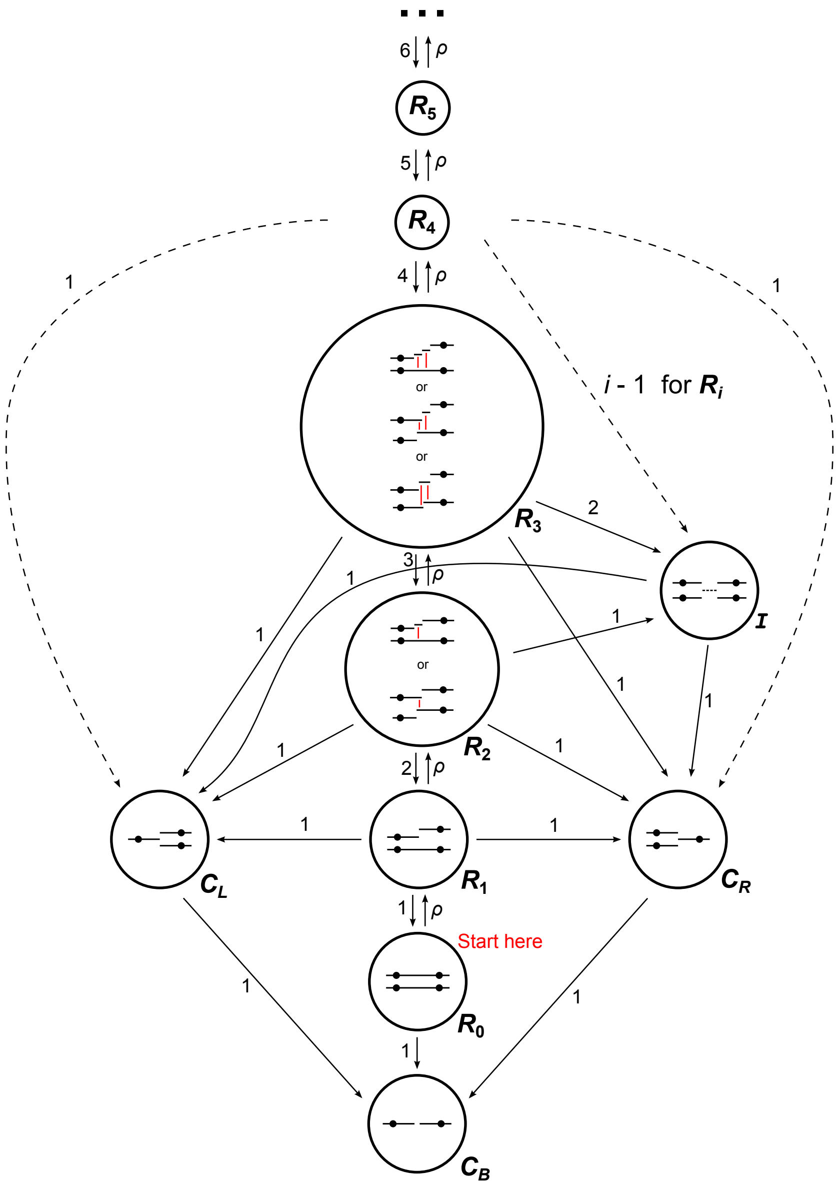

Figure 3.3: Schematic of the SMC' back-in-time Markov jump chain for two loci. Dashed arrows show transition rates that apply for all $\boldsymbol{R_i}$. State $\boldsymbol{I}$ is the state in which some portion of the chromosome between the two focal loci has coalesced but neither focal locus has coalesced. The red lines in states $\boldsymbol{R_2}$ and $\boldsymbol{R_3}$ show the coalescence events that take the process to state $\boldsymbol{I}$.

$i \geq 1)$

$$R_j'(t) = \rho R_{j-1}(t) + (j+1)R_{j+1}(t) - (2j+1+\rho)R_j(t). \tag{3.3}$$

Through substitution, the solution to (3.3) can be shown to be

$$R_j(t) = R_0(t)\frac{\left[\frac{\rho}{2}(1-e^{-2t})\right]^j}{j!}. \tag{3.4}$$

To find $R_0(t)$, we note that it is equal to $f_{T_1,T_2}(t,t)$ (see Eq. (3.2)). In turn,

$$f_{T_1,T_2}(t,t) = f_{T_1}(t)P(T_2 = t|T_1 = t), \tag{3.5}$$

where $f_{T_1}(t) = e^{-t}$ is the marginal distribution of coalescence times at the first (or second) locus and $P(T_2 = t|T_1 = t) = e^{-\rho\lambda(t)}$ is the probability of no change in coalescence times given the coalescence time t at the first locus. Here $\lambda(t) = \frac{1}{4}\left(1 - e^{-2t} + 2t\right)$ is the exponential rate of encountering a change in coalescence time along the chromosome given that the local coalescence time is t [8]. Thus $R_0(t)$ is given by

$$R_0(t) = e^{-t}e^{-\rho\lambda(t)}. \tag{3.6}$$

This completes the solution of $R_j(t)$. Using Figure 3.3,

$$R^+(t) = I(t) + \sum_{j=1}^{\infty} R_j(t), \tag{3.7}$$

where $I(t)$ is the probability that the process is in state $\boldsymbol{I}$ at time t. Using (3.4) and (3.6) we get

$$\begin{aligned}\sum_{j=1}^{\infty} R_j(t) &= R_0(t)\sum_{j=1}^{\infty}\frac{\left[\frac{\rho}{2}(1-e^{-2t})\right]^j}{j!} \\ &= e^{-t}e^{-\frac{\rho}{4}(1+2t-e^{-2t})}\left[e^{\frac{\rho}{2}\left(1-e^{-2t}\right)} - 1\right].\end{aligned} \tag{3.8}$$

Next, $I(t)$ satisfies the forward Kolmogorov equation

$$I'(t) = \sum_{j=2}^{\infty}(j-1)R_j(t) - 2I(t), \tag{3.9}$$

the solution to which is

$$\begin{aligned}
I(t) &= e^{-2t}\int_0^t e^{2u}\sum_{j=2}^{\infty}(j-1)R_j(u)du \\
&= e^{-2t}\int_0^t R_0(u)\left\{2e^{2u} + e^{\frac{\rho}{2}\left(1-e^{-2u}\right)}\left[(\rho-2)e^{2u}-\rho\right]\right\}du \\
&= e^{-2t}\Bigg\{1 - e^{\frac{1}{4}\left(-2t(\rho-2)+\rho-e^{-2t}\rho\right)} \\
&\qquad - e^{-\frac{\rho}{4}}2^{\frac{\rho-4}{2}}(-\rho)^{-\frac{\rho-2}{4}}\left[\Gamma\left(\frac{\rho-2}{4},-\frac{\rho}{4}\right) - \Gamma\left(\frac{\rho-2}{4},-\frac{e^{-2t}\rho}{4}\right)\right]\Bigg\}.
\end{aligned} \tag{3.10}$$

Here, $\Gamma(a,b) = \int_b^{\infty} x^{a-1}e^{-x}dx$ is the incomplete gamma function.

Together (3.6), (3.7), (3.8), and (3.10) give the joint distribution (3.2) for the SMC'. For the ARG and SMC, the expressions for $R_0(t)$ and $R^+(t)$ in the joint distribution (3.2) can be obtained by exponentiating the rate matrices implicit in Figure 3.2. For the SMC, the joint distribution can also be derived using the representation (3.1).

Figure 3.4 compares the joint coalescence time distributions under the SMC and SMC', displaying the differences of these joint distributions with the joint distribution of the ARG. The SMC' provides a much better fit to the ARG joint distribution for the range of recombination rates compared. Both the SMC and the SMC' underestimate the density of outcomes where $T_1 = T_2$, but this underestimation is substantially less under the SMC'.

To summarize the difference between the joint distributions more precisely, we cal-

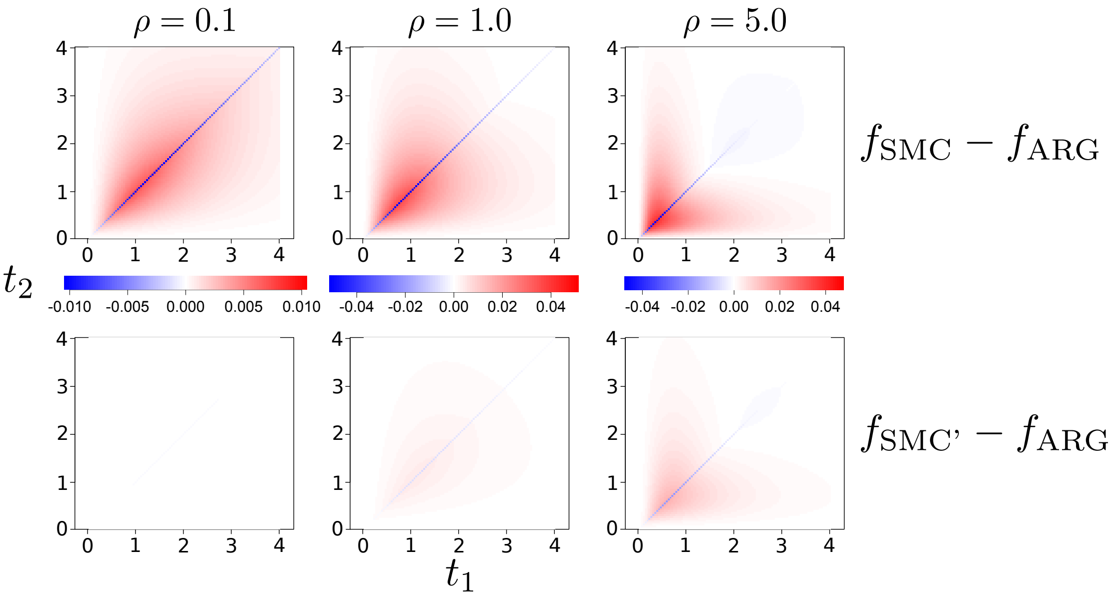

Figure 3.4: Comparison of the difference in joint density of coalescence times $f_{T_1,T_2}(t_1,t_2)$ between the SMC and ARG (top row) and SMC' and ARG (bottom row). Comparisons are made for three different rates of recombination between the two focal loci ($\rho = 0.1, 1.0, 5.0$).

culated the total variation distance between the SMC and ARG and between the SMC' and ARG across a range of recombination rates. The total variation distance between the SMC and the ARG is defined as

$$TV\,(\text{SMC}, \text{ARG}) = \frac{1}{2}\int_0^\infty \int_0^\infty \left| f^{\text{SMC}}(t_1,t_2) - f^{\text{ARG}}(t_1,t_2)) \right| dt_2 dt_1, \tag{3.11}$$

where $f^{\text{SMC}}(t_1,t_2)$ and $f^{\text{ARG}}(t_1,t_2)$ are the joint densities $f_{T_1,T_2}(t_1,t_2)$ defined under the SMC and ARG, respectively. The total variation distance between the SMC' and ARG is similarly defined. Figure 3.5 shows the total variation distance from the ARG for the SMC and SMC' over a range of recombination rates. Total variation distances were calculated numerically. For both the SMC and SMC', the total variation distance was maximized at some intermediate recombination rate, approximately $\rho = 1.1$ for the SMC and $\rho = 3.2$ for the SMC'.

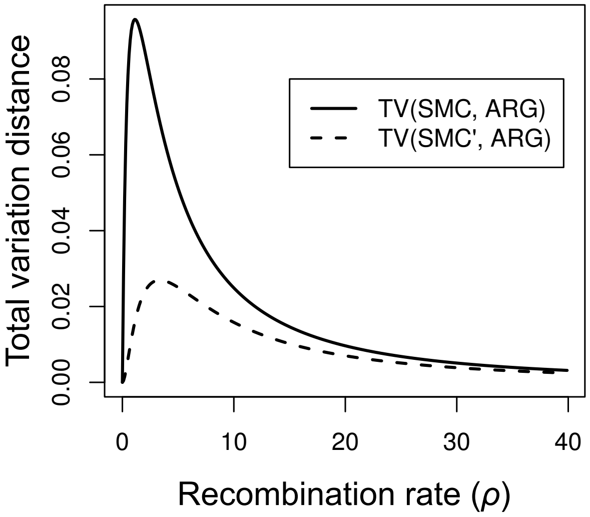

Figure 3.5: Total variation distance between the SMC and ARG (solid line) and the SMC' and ARG (dashed line) as a function of recombination rate. Total variation distances were calculated numerically.

It is interesting to note that the walk on the states $\boldsymbol{R_0}$, $\boldsymbol{R_1}$, $\boldsymbol{R_2}$, ..., constitutes a birth-death process with killing, where birth events correspond to additional recombination events taking the process from $\boldsymbol{R_i}$ to $\boldsymbol{R_{i+1}}$, death events correspond to coalescence events that take the process from $\boldsymbol{R_i}$ to $\boldsymbol{R_{i-1}}$, and killing events, which take the process to an absorbing state, here correspond to coalescence events that take the process to $\boldsymbol{C_L}$, $\boldsymbol{C_R}$, or $\boldsymbol{I}$. Under this formulation, the birth rate is constant $\lambda_i = \rho$, the death rate is linear $\mu_i = i$, and the killing rate is linear $\gamma_i = i + 1$. This class of processes was studied by van Doorn and Zeifman [60], who demonstrated a different approach for calculating $R_i(t)$. This alternative approach (not shown) confirms our derivation of

(3.6).

3.2.3 Conditional distribution of coalescence times

In this section we consider the distribution of coalescence times at one locus given the coalescence time at the other. The conditional density of T_2 given T_1, $f_{T_2|T_1}(t_2|t_1)$, can be calculated by dividing the joint distribution $f_{T_1,T_2}(t_2, t_1)$ by e^{-t_1}, the marginal distribution of coalescence times at the left locus:

$$f_{T_2|T_1}(t_2|t_1) = \frac{f_{T_1,T_2}(t_1, t_2)}{e^{-t_1}}. \tag{3.12}$$

Hobolth and Jensen [59] introduced a framework for modeling the distribution of T_2 given T_1 using a time-inhomogeneous continuous-time Markov chain. (Note that the model called SMC' in Hobolth and Jensen [59] is an SMC'-like model of two loci that is not based on the continuous-chromosome SMC'. It is different from the SMC' model we consider here.) This framework can be extended to the SMC', producing the continuous-time Markov chain shown in Figure 3.6. Within this framework, a coalescence time T_2 at the right locus is generated back in time conditioned upon a coalescence time T_1 at the left locus.

Figure 3.7 compares the conditional density $f_{T_2|T_1}(t_2|t_1)$ of coalescence times t_2 at the right locus conditioned upon the coalescence times t_1 at the left locus for different values of t_1 and recombination rate ρ. The conditional density under the SMC' is much closer to the density produced by the ARG than is the conditional density under the SMC.

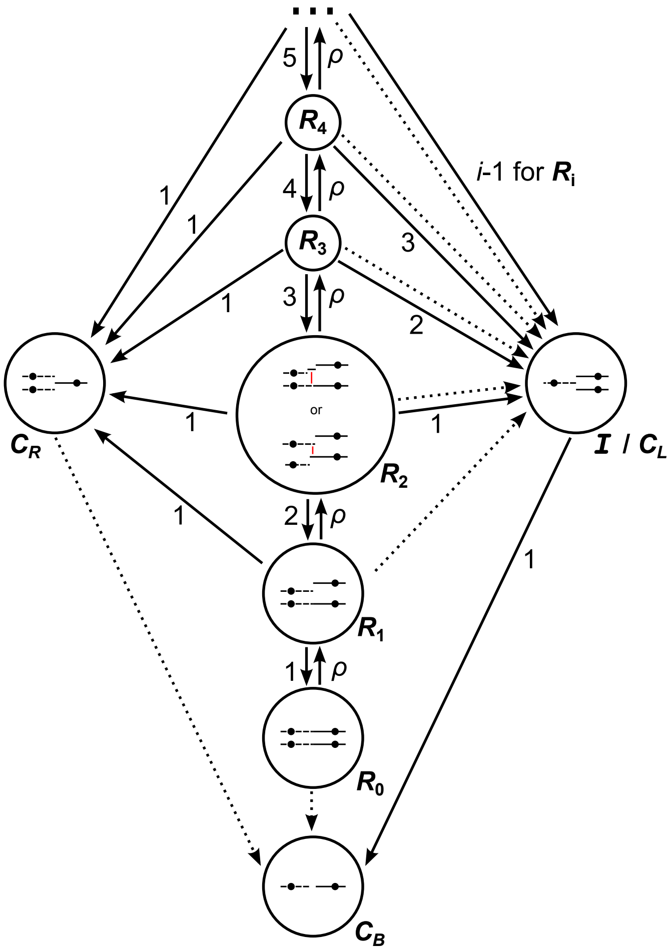

Figure 3.6: Back-in-time Markov jump chain for generating a coalescence time T_2 at the right locus conditional on the time $T_1 = t_1$ at the left locus under the SMC'. Starting at time zero in state $\boldsymbol{R_0}$, the process follows the transitions indicated by the solid arrows at the rates accompanying these arrows. Transitions indicated by dotted arrows are followed instantaneously at time t_1. See Hobolth and Jensen [59] for analogous processes for the ARG and SMC models.

3.2.4 Probability of coalescence times being equal

In the two-locus, back-in-time ancestral models, T_1 and T_2 are equal when the state $\boldsymbol{C_B}$ is entered through $\boldsymbol{R_0}$ rather than $\boldsymbol{C_L}$ or $\boldsymbol{C_R}$. That is, the coalescence times are equal when a coalescence event occurs between two ancestral lineages each carrying ancestral material at both of the focal loci. The probability $P(T_1 = T_2)$ can be obtained by

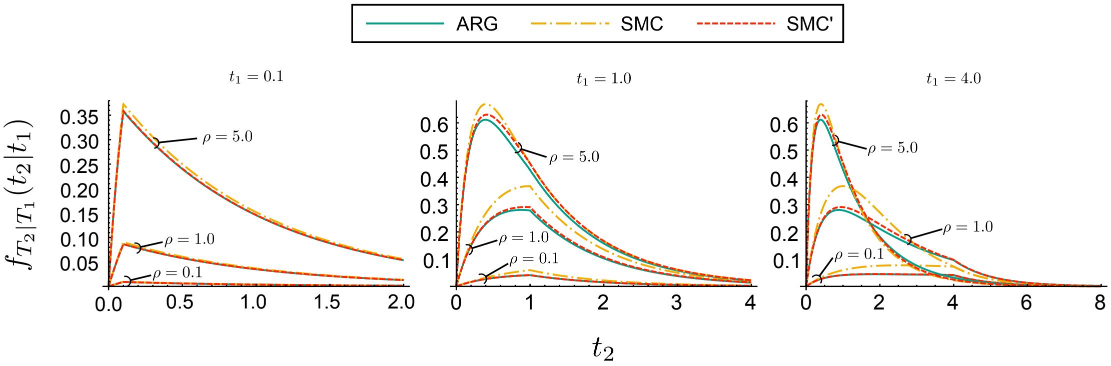

Figure 3.7: Comparison of densities of coalescence times t_2 at the right locus conditioned upon coalescence times t_1 at the left locus. Conditional densities $f_{T_2|T_1}(t_2|t_1)$ are shown for the ARG, SMC, and SMC' models for three different rates of recombination between the two loci ($\rho = 0.1, 1.0, 5.0$) and three different conditioned-upon coalescence times t_1 at the left locus ($t_1 = 0.1, 1.0, 4.0$). The area under each curve is $P(T_2 \neq t_1|T_1 = t_1)$; the conditional probabilities $P(T_2 = t_1|T_1 = t_1)$ are not shown.

analyzing the ancestral processes introduced in Section 3.2.1. For the SMC and SMC', $P(T_1 = T_2)$ can also be obtained by considering the original, sequential formulations of McVean and Cardin [10] and Marjoram and Wall [11], respectively.

For the ARG, Simonsen and Churchill [58] showed that the probability that T_1 is equal to T_2 is

$$P_{\mathrm{ARG}}(T_1 = T_2) = \frac{\rho + 18}{\rho^2 + 13\rho + 18}. \tag{3.13}$$

Under the SMC, representation (3.1) shows that

$$P_{\mathrm{SMC}}(T_1 = T_2) = \frac{1}{1 + \rho}. \tag{3.14}$$

Under the SMC', the probability that T_1 is equal to T_2 is most easily obtained by considering the sequential formulation of Carmi et al. [8]. Under this formulation, when the local coalescence time is t, the distance until a change in the coalescence time (measured in units of the scaled recombination parameter ρ) is exponentially distributed with rate parameter $\lambda(t) = \frac{1}{4}\left(1 - e^{-2t} + 2t\right)$, as above. Thus

$$\begin{aligned} P_{\text{SMC'}}(T_1 = T_2) &= \int_0^\infty e^{-t} e^{-\rho\lambda(t)} dt \\ &= 2^{\rho/2} e^{-\rho/4} (-\rho)^{-\frac{1}{2}-\frac{\rho}{4}} \left[\Gamma\left(\frac{2+\rho}{4}\right) - \Gamma\left(\frac{2+\rho}{4}, -\frac{\rho}{4}\right) \right]. \end{aligned} \tag{3.15}$$

The first line in (3.15) follows from the fact that the marginal distribution of coalescence times at the left locus is e^{-t}, and given the left coalescence time t, the probability that the same coalescence time extends a distance at least ρ to the right is $e^{-\rho\lambda(t)}$. This equation was also derived by Eriksson et al. [61] using a similar approach (see their Eq. (10)). Figure 3.8 compares $P(T_1 = T_2)$ for the ARG, SMC, and SMC'.

3.2.5 Covariance of coalescence times

The covariance between coalescence times, $\text{Cov}[T_1, T_2]$, is a measure of the dependence between coalescence times and is informative about the scale over which features of the genome become independent. Under the ARG,

$$\text{Cov}[T_1, T_2] = \frac{\rho + 18}{\rho^2 + 13\rho + 18} \tag{3.16}$$

[62, 63]. Under the SMC,

$$\text{Cov}[T_1, T_2] = \frac{1}{1 + \rho} \tag{3.17}$$

[10].

For both the ARG and the SMC, $P(T_1 = T_2)$ is equal to $\text{Cov}[T_1, T_2]$. This can be shown to hold in general for two-locus coalescence models where the marginal distribution of coalescence times is exponential with rate 1. Using the definition of covariance:

$$\begin{aligned} \operatorname{Cov}[T_1, T_2] &= \mathrm{E}[T_1 T_2] - \mathrm{E}[T_1]\,\mathrm{E}[T_2] \\ &= \mathrm{E}[T_1 T_2] - 1. \end{aligned} \tag{3.18}$$

The expectation $\mathrm{E}[T_1 T_2]$ can be derived using the fact that $(a-b)^2 = a^2 + b^2 - 2ab$:

$$\begin{aligned} 2\,\mathrm{E}[T_1 T_2] &= \mathrm{E}[T_1^2] + \mathrm{E}[T_2^2] - \mathrm{E}[(T_1 - T_2)^2] \\ &= 2 + 2 - E[(T_1 - T_2)^2 | T_1 \neq T_2] P(T_1 \neq T_2) \\ &= 4 - 2P(T_1 \neq T_2). \end{aligned} \tag{3.19}$$

The final equality in (3.19) follows from the fact that $|T_1 - T_2|$ has an exponential distribution with rate 1 when $T_1 \neq T_2$. Therefore $\mathrm{E}[T_1 T_2] = 2 - P(T_1 \neq T_2)$ and

$$\begin{aligned} \operatorname{Cov}[T_1, T_2] &= \mathrm{E}[T_1 T_2] - 1 \\ &= 2 - P(T_1 \neq T_2) - 1 \\ &= 1 - P(T_1 \neq T_2) \\ &= P(T_1 = T_2). \end{aligned} \tag{3.20}$$

Thus Figure 3.8 compares both $P(T_1 = T_2)$ and $\operatorname{Cov}[T_1, T_2]$ under the ARG, SMC, and SMC'.

Ohta and Kimura [64] introduced the approximation

$$\sigma_d^2 = \frac{\mathrm{E}[D]}{\mathrm{E}[p_1(1-p_1)p_2(1-p_2)]} \tag{3.21}$$

of the linkage disequilibrium measure

$$r^2 = \mathrm{E}\left[\frac{D}{p_1(1-p_1)p_2(1-p_2)}\right], \tag{3.22}$$

where $D = p_{12} - p_1 p_2$ is the standard measure of linkage disequilibrium at two partially linked loci, p_1 and p_2 are allele frequencies at the two loci, and p_{12} is the frequency of gametes carrying both of the alleles represented by p_1 and p_2. McVean [65] showed that σ_d^2 could be expressed in terms of the covariances of coalescence times:

$$\sigma_d^2 = \frac{C_{ij,ij} - 2C_{ij,ik} + C_{ij_kl}}{C_{ij,kl} + 1}. \tag{3.23}$$

Here $C_{ij,kl} = \mathrm{Cov}[T_1^{(ij)}, T_2^{(kl)}]$ is the covariance of the coalescence time at the first locus sampled from haplotypes i and j and the coalescence time at the second locus sampled from haplotypes k and l. The above proof that $P(T_1 = T_2) = \mathrm{Cov}[T_1, T_2]$ applies to $C_{ij,kl}$ as well, regardless of whether i is the same as k or j is the same as l, since the marginal coalescence time at a single locus is still exponentially distributed with rate 1 regardless of the initial configuration. We were unable to solve for these probabilities under both the SMC and SMC'.

In order to calculate the necessary covariances, McVean and Cardin [10] used the simplifying assumption that all recombination occurred at the same point between the two loci. We note that this assumption makes the SMC' equivalent to the ARG. Marjoram and Wall [11] simulated mean values of r^2 at different genomic distances under the ARG, SMC', and SMC, showing that the SMC' is more similar to the ARG in mean r^2 values than was the SMC.

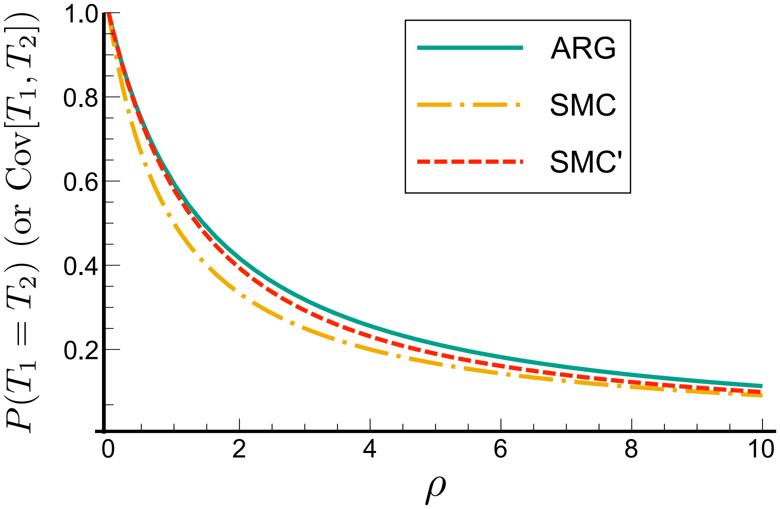

Figure 3.8: Comparison of $P(T_1 = T_2)$ for the ARG, SMC, and SMC'. For these three models, this probability is equal to $\mathrm{Cov}[T_1, T_2]$ (see text).

Covariance of coalescence times when ρ is small

It is interesting to consider $\mathrm{Cov}[T_1, T_2] = P(T_1 = T_2)$ when ρ is small. For the ARG, consideration of (3.13) shows that $\mathrm{Cov}[T_1, T_2] = P_{\mathrm{ARG}}(T_1 = T_2) = 1 - 2\rho/3 + O(\rho^2)$. Likewise, for the SMC, (3.14) shows that $\mathrm{Cov}[T_1, T_2] = P_{\mathrm{SMC}}(T_1 = T_2) = 1 - \rho + O(\rho^2)$. For the SMC', the integral representation of $P_{\mathrm{SMC'}}(T_1 = T_2)$ in (3.15) allows for the calculation of this quantity as a first-order expansion in ρ:

$$
\begin{aligned}
\mathrm{Cov}[T_1, T_2] = P_{\mathrm{SMC'}}(T_1 = T_2) &= \int_0^\infty e^{-t} e^{-\rho\lambda(t)} dt \\
&= \int_0^\infty e^{-t} \left[1 - \lambda(t)\rho + O(\rho^2)\right] dt \\
&= 1 - \rho \int_0^\infty e^{-t} \lambda(t) dt + O(\rho^2) \\
&= 1 - \frac{2\rho}{3} + O(\rho^2).
\end{aligned}
\tag{3.24}
$$

Thus, $\mathrm{Cov}[T_1, T_2]$ (or $P(T_1 = T_2)$) is the same up to order ρ^2 under the ARG and SMC'.

3.2.6 Coalescence times at recombination sites

In this section, we show that the joint distribution of coalescence times on either side of a recombination event is the same under the SMC' and marginally under the ARG, and we derive this distribution. Consider the continuous-time Markov chains representing the two-locus ARG and SMC' models (Figs. 3.2A and 3.3, respectively) in the limit $\rho \to 0$ and conditioning on the first event being a recombination event. These models represent the joint distribution of coalescence times on either side of a recombination event under the ARG and SMC'. In both of these conditional continuous-time Markov chains, the waiting time until the first event, conditional on that event being a recombination event, has an exponential distribution with rate $1 + \rho$, which converges to 1 as $\rho \to 0$. After that first recombination event, the rate of all additional recombination events converges to zero in the $\rho \to 0$ limit, so all of the remaining events must be coalescence events, each of which occurs with rate 1. Under the ARG and the SMC', the coalescence events that are possible from state $\boldsymbol{R_1}$ are the same. Thus, the joint distribution of coalescence times at recombination sites is the same under the SMC' and the ARG.

Figure 3.9A shows the two-locus continuous-time Markov chain representing this conditional process representing coalescence times on either side of a recombination event under the the ARG and SMC'. This Markov chain starts in a special initial state $\boldsymbol{R_0^*}$, out of which the first event is always a recombination event, which happens with rate 1, as described above. In previous sections, we used T_1 and T_2 to represent the coalescence times at two loci some fixed distance apart. To avoid confusion, in this section we use S and T to represent the coalescence times on the left and right sides of a recombination event, respectively.

Recombination events are visible in sequence data only if they change the local coalescence time. Thus, it is of special interest to condition on $S \neq T$ in the above model in order to derive the joint distribution of coalescence times on either side of a change in coalescence times under the ARG and SMC'. Conditioning on $S \neq T$, the transition out of $\boldsymbol{R_1}$ must be into either $\boldsymbol{C_L}$ or $\boldsymbol{C_R}$. These transitions occur with conditional rate 3/2, since the total rate of leaving $\boldsymbol{R_1}$ is three in the unconditional model, and two of the ways of leaving $\boldsymbol{R_1}$ result in the coalescence times being different.

The model representing coalescence times on either side of a change in coalescence times (i.e., where at recombination sites where $S \neq T$) is shown in Figure 3.9B. Under this model, the joint distribution of S and T is that of

$$(S, T) \sim \big(X_1 + X_2 + RX_3,\ X_1 + X_2 + (1 - R)X_3\big), \tag{3.25}$$

where $X_1 \sim \text{Exp}(1)$, $X_2 \sim \text{Exp}(3)$, $R \sim \text{Bernoulli}(1/2)$, $X_3 \sim \text{Exp}(1)$, and each random variable is independently distributed. The joint density function of S and T under this model is

$$f_{S,T}(s,t) = \begin{cases} \frac{3}{4}\left(1 - e^{-2s}\right) e^{-t} & s < t \\ \\ \frac{3}{4}\left(1 - e^{-2t}\right) e^{-s} & s > t, \end{cases} \tag{3.26}$$

and the marginal density function of S (or T) is

$$\pi(s) = \frac{3}{8} e^{-s}\left(2s + 1 - e^{-2s}\right). \tag{3.27}$$

The conditional distribution of T given S is

$$f_{T|S}(t|s) = \frac{f_{S,T}(s,t)}{\pi(s)} = \begin{cases} \frac{2\left(1-e^{-2t}\right)}{1-e^{-2s}+2s} & t < s \\ \\ \frac{2e^{-(t-s)}\left(1-e^{-2s}\right)}{1-e^{-2s}+2s} & t > s. \end{cases} \tag{3.28}$$

Equations (3.26), (3.27), and (3.28) hold marginally at recombination sites where the coalescence time changes under both the ARG and SMC'. Equations (3.27) and (3.28) were derived for the SMC' by Carmi et al. [8, see eqns. (8) and (9), respectively], confirming our derivation here.

Note that the model representing the joint distribution of coalescence times at recombination sites under the SMC is equivalent to the model in Figure 3.9B with the transition rates from $\boldsymbol{R_1}$ to $\boldsymbol{C_L}$ and $\boldsymbol{C_R}$ equal to 1 instead of 3/2. Under this model for the SMC, the joint distribution of coalescence times on either side of a recombination event is that of

$$(S,T) \sim (X_1 + X_2, X_1 + X_3), \tag{3.29}$$

where X_1, X_2, and X_3 are all mutually independent exponential random variables with rate 1. The joint density of S and T under the SMC is

$$f_{S,T}(s,t) = \begin{cases} e^{-t}(1-e^{-s}) & s < t \\ \\ e^{-s}(1-e^{-t}) & s > t \end{cases} \tag{3.30}$$

and the marginal density of S (or T) is

$$\pi(s) = se^{-s}. \tag{3.31}$$

Under the SMC, the conditional distribution of T given S is

$$f_{T|S}(t|s) = \frac{f_{S,T}(s,t)}{\pi(s)} = \begin{cases} \frac{1-e^{-t}}{s} & t < s \\ \frac{e^{-(t-s)}(1-e^{-s})}{s} & t > s, \end{cases} \tag{3.32}$$

which confirms the derivation of Li and Durbin [12, cf. their Eq. (S6)].

SMC' as canonical first-order Markov approximation to ARG

Under the sequential formulation of each model considered here, the infinitesimal probability of a recombination event occurring in the interval $(x, x+dx)$ given the coalescence time s at x is $s\,dx$. This fact, together with the fact that the joint distribution of coalescence times at recombination sites is the same under the ARG and SMC' (whether or not the coalescence time changes), implies that the conditional distribution of coalescence times at point $x + dx$ given the coalescence time at point x is the same under the SMC' and ARG.

This result demonstrates that the pairwise SMC' is the canonical first-order Markov approximation to the pairwise ARG. Given an infinite-order Markov chain $\{X_i, i = 0, 1, 2, ...\}$, where the distribution of each X_j depends on all previous X_i, $i < j$, the canonical k-order Markov approximation to $\{X_i\}$ is the Markov chain $\{X_i^{[k]}\}$ satisfying

$$P(X_n^{[k]}|X_{n-1}^{[k]} = x_{n-1}, \ldots, X_{n-k}^{[k]} = x_{n-k}) = P(X_n|X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}). \tag{3.33}$$

That is, the transition probabilities under the k-order canonical Markov approximation are equal to the transition probabilities conditional on the previous k states under the infinite-order chain. See Schwarz [66], Fernández and Galves [67], and Gallo et al. [68] for examples of mathematical studies of canonical Markov approximations of infinite-order Markov chains.

Here we informally extend the terminology of canonical Markov approximations to continuous processes. The SMC' is the canonical first-order Markov approximation to the ARG because the distribution of coalescence times at $x + dx$ conditional on the coalescence time at x is the same under the ARG (an infinite-order, sequentially non-Markovian continuous process) and the SMC' (a first-order sequentially Markov continuous process).

3.2.7 Asymptotic bias of the population-size estimators under SMC and SMC'

Given the joint density of pairwise coalescence times at recombination sites under the ARG, it is possible to determine the asymptotic bias of maximum-likelihood estimators of population size derived from the pairwise SMC and SMC' likelihood functions. These likelihood functions give the probability of observing a sequence of pairwise coalescence

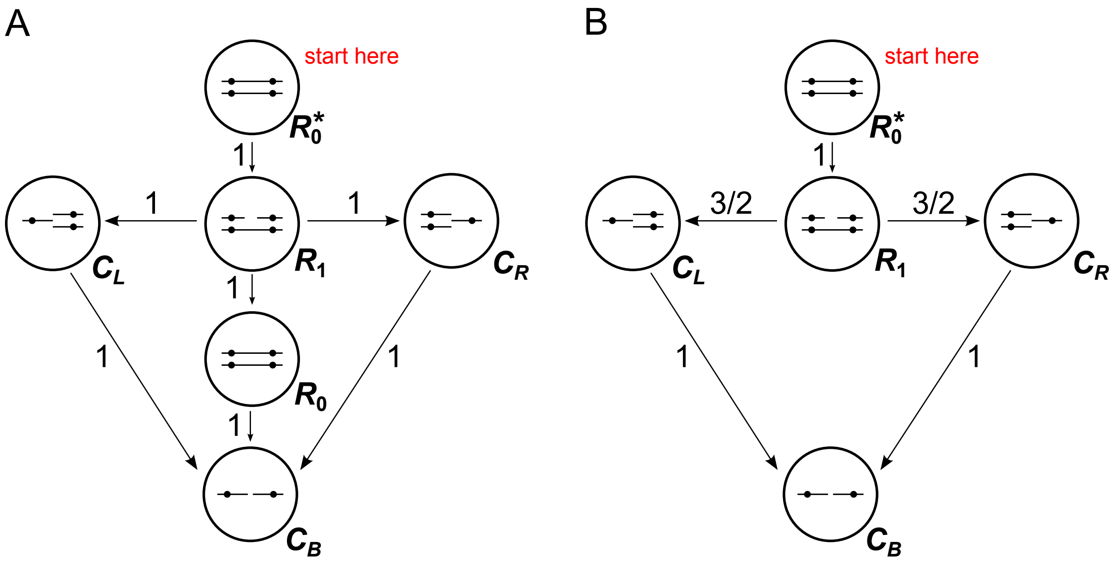

Figure 3.9: Two-locus continuous-time Markov chains representing the ARG and SMC' models in the $\rho \to 0$ limit, conditional on the first event being a recombination event. These models represent the joint distribution of coalescence times at recombination sites in the ARG and SMC'. The state $\boldsymbol{R_0^*}$ represents a special starting state out of which the first event is always a recombination event. Panel A shows the process unconditional on whether $S = T$, and Panel B shows the process conditional on $S \neq T$. The model representing the joint distribution of coalescence times at recombination sites under the SMC is equivalent to the model in Panel B with the transition rates from $\boldsymbol{R_1}$ to $\boldsymbol{C_L}$ and $\boldsymbol{C_R}$ equal to 1 instead of $3/2$.

times and corresponding segment lengths across a chromosome under the SMC and SMC' models. Related likelihood functions (allowing for variable historical population size) are implicitly maximized in the PSMC and MSMC inference procedures [12, 17, respectively]. These inference procedures are hidden Markov model (HMM) methods in which the local coalescence times (or genealogies) and segment lengths are hidden states inferred from sequence data. Here, we consider the estimators that would be obtained if the hidden states in these models were actually observable [see also 69]. We are motivated by the fact that any properties of the estimators we consider here are likely to be properties of the full HMM-based inference procedures.

To investigate the asymptotic properties of these estimators, we assume that data are generated under the ARG, such that at a fixed point the distribution of pairwise coalescence times is exponential with rate equal to 1 and an ancestral segment of length l recombines back in time at rate $\rho l/2$. Here, segment lengths are measured in units of the true scaled recombination parameter ρ. Data generated under this model can be represented as a sequence of pairwise coalescence times and corresponding segment lengths: $\{(t_i, l_i) : 1 \leq i \leq k\}$.

We are interested in estimating a single relative population size η (defined relative to the true population size, N), which must be incorporated into the transition density function $q(t|s)$ at recombination sites under the SMC and SMC'. Under the SMC, this transition density function is

$$q_{\mathrm{SMC}}(t|s;\eta) = \begin{cases} \frac{1}{s}(1 - e^{-t/\eta}) & t < s \\ \\ \frac{1}{s}e^{-(t-s)/\eta}(1 - e^{-s/\eta}) & t > s. \end{cases} \tag{3.34}$$

This is equivalent to the conditional density (3.32) above with the addition of a relative population size parameter. Under the SMC', the transition function is

$$q_{\mathrm{SMC'}}(t|s;\eta) = \begin{cases} \frac{\frac{2}{\eta}\left(1-e^{-2t/\eta}\right)}{1+\frac{2s}{\eta}-e^{-2s}} & t < s \\ \\ \frac{\frac{2}{\eta}e^{-(t-s)/\eta}\left(1-e^{-2s/\eta}\right)}{1+\frac{2s}{\eta}-e^{-2s}} & t < s, \end{cases} \tag{3.35}$$

which is equivalent to the conditional density (3.28) with a relative population size parameter included.

Under the SMC, given the local coalescence time t, the distance along the chromosome until the nearest recombination event (measured in units of ρ) is exponentially distributed with rate t [10]. The likelihood function for a single relative population size

η under the SMC is thus

$$\begin{aligned} L_{\text{SMC}}(\eta|\{(t_i, l_i)\}) &= \frac{1}{\eta} e^{-\frac{t_1}{\eta}} \prod_{i=2}^{k} q_{\text{SMC}}(t_i|t_{i-1};\eta) \prod_{i=1}^{k} t_i e^{-t_i l_i} \\ &\propto \frac{1}{\eta} e^{-\frac{t_1}{\eta}} \prod_{i=2}^{k} q_{\text{SMC}}(t_i|t_{i-1};\eta). \end{aligned} \tag{3.36}$$

Under the SMC', the likelihood function for a relative population size η is

$$L_{\text{SMC}'}(\eta|\{(t_i, l_i)\}) = \frac{1}{\eta} e^{-\frac{t_1}{\eta}} \prod_{i=2}^{k} q_{\text{SMC}'}(t_i|t_{i-1};\eta) \prod_{i=1}^{k} \lambda(t_i, \eta) e^{-\lambda(t_i,\eta) l_i}, \tag{3.37}$$

where $\lambda(t, \eta) = \frac{1}{4}\left[\eta(1 - e^{-2t/\eta}) + 2t\right]$ is the exponential rate of encountering recombination events that change the coalescence time when the local coalescence time is t (see above and Carmi et al. [8]). Note that under the SMC, the length l_i of a segment is independent of the relative population size η given the local coalescence time t_i. This is not true for the SMC', since the probability that the coalescence time changes at a recombination site depends on the population size.

For a given set of observations $\{(t_i, l_i)\}$, the maximum-likelihood estimate $\hat{\eta}$ of the relative population size under the SMC is

$$\hat{\eta} = \underset{\eta}{\text{argmax}}\ L(\eta|\{(t_i, l_i)\}) = \underset{\eta}{\text{argmax}}\ \frac{1}{\eta} e^{-\frac{t_1}{\eta}} \prod_{i=2}^{k} q_{\text{SMC}}(t_i|t_{i-1};\eta). \tag{3.38}$$

As the length of the chromosome increases and the number of coalescence-time changes goes to infinity, the asymptotic maximum-likelihood estimate $\hat{\eta}^*$ of the relative population size under the SMC is

$$\begin{aligned}
\hat{\eta}^* &= \lim_{k\to\infty} \underset{\eta}{\operatorname{argmax}} \ \frac{1}{\eta} e^{-\frac{t_1}{\eta}} \prod_{i=2}^{k} q_{\mathrm{SMC}}(t_i|t_{i-1};\eta) \\
&= \lim_{k\to\infty} \underset{\eta}{\operatorname{argmax}} \ \left\{ \log\left(\frac{1}{\eta} e^{-\frac{t_1}{\eta}}\right) + \sum_{i=2}^{k} \log\left[q_{\mathrm{SMC}}(t_i|t_{i-1};\eta)\right] \right\} \\
&= \lim_{k\to\infty} \underset{\eta}{\operatorname{argmax}} \ \sum_{i=2}^{k} \log(q_{\mathrm{SMC}}(t_i|t_{i-1};\eta)) \\
&= \underset{\eta}{\operatorname{argmax}} \ \mathrm{E}_{\mathrm{ARG}}\big[\log(q_{\mathrm{SMC}}(T|S;\eta))\big] \\
&= \underset{\eta}{\operatorname{argmax}} \ \int_0^\infty \int_0^\infty \pi_{\mathrm{SMC}'}(s) q_{\mathrm{SMC}'}(t|s;1) \log\left(q_{\mathrm{SMC}}(t|s;\eta)\right) dtds \\
&\approx 0.95.
\end{aligned} \tag{3.39}$$

Here the penultimate equality holds only if there is ergodic (i.e., law-of-large-numbers-like) convergence of the sequence of pairs of coalescence times on either side of a recombination site under the ARG. In Appendix C, we show that the continuous-chromosome pairwise ARG is ergodic. That is, the mean coalescence time across a long chromosome converges to the mean coalescence time at a single point along the chromosome. We are unable to prove the ergodicity of the sequence of pairs of coalescence times at recombination sites where the coalescence time changes; instead, we note that (3.39) is supported by simulation (see below). We also note that Wiuf [70] proved the ergodicity of the discrete-locus ARG under a variety of neutral demographic models. A similarly in-depth proof may also apply to (3.39) in the context of continuous-chromosome models, but we do not explore the point further.

In (3.39), the ultimate equality follows from the fact that the joint distribution of coalescence times is marginally the same at recombination sites under the ARG and the SMC'. Numerical maximization of the double integral shows that the maximum-likelihood estimate of a single population size N under the pairwise SMC is asymptotically biased, with the asymptotic estimate being approximately $0.95N$.

Under the ARG, the stationary distribution of lengths between recombination events that change the local coalescence time (i.e., the identity-by-descent segment length distribution) is slightly different from that of the SMC'. (They are different because subsequent recombination events "heal" with slightly different probabilities under the ARG, while under the SMC', each subsequent recombination event heals with the same probability.) Under the ARG, the identity-by-descent (IBD) length distribution is not currently known. Given that under the SMC' the maximum-likelihood estimator for a relative population size involves the observed lengths, it is not currently possible to calculate the asymptotic bias of the pairwise SMC' maximum-likelihood estimator of a single population size. However, the IBD length distribution under the ARG is approximated very closely by the SMC' IBD length distribution [8], so the SMC' estimator is likely to be nearly asymptotically unbiased.

We propose the following estimator, which should be asymptotically unbiased for data generated by the ARG:

$$\hat{\eta}' = \operatorname*{argmax}_{\eta} \; \frac{1}{\eta} e^{-\frac{t_1}{\eta}} \prod_{i=2}^{k} q_{\mathrm{SMC}'}(t_i | t_{i-1}; \eta). \tag{3.40}$$

This estimator is unbiased under the same assumption that was used to calculate the asymptotic bias of the SMC above, which is that the sequence of pairs of coalescence times are ergodic across an infinitely long chromosome.

We confirm the asymptotic bias of the SMC estimator and the apparent lack of asymptotic bias of the SMC' estimator and $\hat{\eta}'$ by simulation. Figure 3.10 shows 100 simulated estimates calculated using the SMC, SMC', and SMC'-lengths-only likelihood functions. Each estimate was calculated using 100 independent pairs of chromosomes simulated under the ARG, with each chromosome of total length $4Nl = 1000$, where N is the diploid size and l is the length in Morgans. To calculate these estimators for multiple chromosomes, all likelihood functions were multiplied across independent pairs of chromosomes. The same set of simulations was used to produce the estimators for all three likelihood functions.

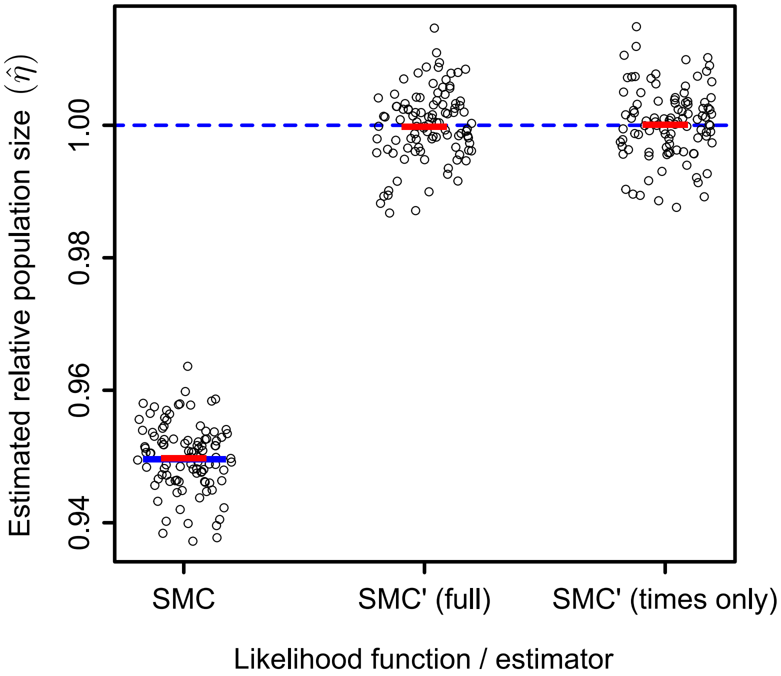

Figure 3.10: Maximum-likelihood estimates of relative population size with three different Markov chain likelihood functions. For each simulation, the segment lengths and coalescence times were taken from 100 independent pairs of chromosomes, with each chromosome being of length $\rho = 4Nr = 1000$ simulated under the ARG. A maximum-likelihood estimate was calculated using the SMC, SMC', and times-only SMC' likelihood functions (equations (3.36), (3.37), and (3.40), respectively). The true scaled population size is $\eta = 1$, shown with the dashed blue line. The predicted asymptotic bias of the SMC likelihood function ($\hat{\eta} = 0.95$) is shown with a solid blue line. The sample mean of the estimates calculated with each likelihood function is shown with a solid red line. A total of 100 simulated datasets were analyzed.

3.3 Discussion

We have proposed a model that describes the pairwise coalescence times at two fixed loci evolving under the SMC' model of coalescence with recombination. We analyzed this model to derive quantities that have not been derived previously for the SMC', including the joint density of coalescence times (unconditional and conditional), the probability that the coalescence times are the same, and the covariance of the two coalescence times, which was shown to be equal to the probability that the coalescence times are the same for the ARG, SMC, and SMC'. We compared these quantities against those produced by the ARG and SMC models. In every comparison, the difference between the ARG and the SMC' was much less than the difference between the ARG and the SMC.

We also showed that the conditional distribution of coalescence times at point $x + dx$ given the coalescence time at x is the same under the ARG and SMC'. This implies that the SMC' is the canonical first-order approximation to the pairwise ARG. However, this correspondence is true only of the continuous-chromosome models. If instead the ARG is a model of the genealogies at a sequence of discrete loci, then the first-order canonical Markov approximation is the Markov approximation obtained by modeling a conditional ARG between every successive pair of loci. This model was studied by Hobolth and Jensen [59], who referred to the model as a "natural" Markov approximation to the ARG. Chen et al. [71] presented a method of simulating data under higher-order sequentially Markov approximations to the ARG, where the ARG of some number of preceding loci is retained in the process of generating the marginal genealogy at a given locus. They showed by simulation that higher-order approximations generate times until most recent common ancestry that are more consistent with the ARG than do lower-order approximations, but little theoretical work on these higher-order Markov approximations has been done.

We showed that the maximum-likelihood estimate of a single population size under the pairwise SMC is asymptotically biased, producing an estimate of about 95% of the true population size. Given the current widespread use of the SMC model in population-genomic inference methods [12, 13, 14, 16, 56], there is an apparent need to re-examine the consequences of using the simpler SMC model instead of the slightly more complicated SMC' model. For example, it will be important to consider whether including the possibility of varying population sizes, as for example is done in the PSMC HMM inference method [12], will increase or decrease asymptotic bias. In this context, using the SMC as a basis for a likelihood function may also bias the estimates of the timing of population size changes, since the longer segments produced by the ARG will seem younger when they are modeled under the SMC.

From the arguments that led to the development of the models in Figure 3.9, it seems that variable population size or population substructure will not change the fact that the joint distribution of coalescence times at recombination sites is the same under the SMC' and marginally under the ARG. Changing the population size to a function $\eta(t)$ and the recombination rate to a function $\rho(t) = \rho_0 \eta(t)$ does not change the previous arguments, so long as $\rho(t) \to 0$ as the distance ρ_0 between the two loci goes to zero. For example, regardless of population size, the waiting time until the conditioned-upon recombination event will be the same under the SMC' and ARG, and the remaining coalescence events would always be distributed identically between the SMC' and ARG. Similarly, when there are more than two haplotypes sampled, it seems that the joint distributions of genealogies on either side of a recombination event would be the same between the SMC' and the ARG marginally. These ideas need to be properly explored in future studies.

The SMC' is the model underlying two recently introduced population-genetic inference methods: the multiple SMC (MSMC) method of Schiffels and Durbin [17] (which simplifies to a PSMC' inference procedure when the number of haplotypes is two) and a procedure based on the distribution of distances between heterozygous bases, introduced by Harris and Nielsen [52]. In each case it was acknowledged that the SMC' provided more accurate results than the SMC. In light of the results we present here, we suggest that whenever a first-order sequentially Markov coalescent model is needed, the SMC' should be used whenever the calculations are possible.

4

Inference of demographic and reproductive history using a triploid sequentially Markov coalescent model

One of the most persistent questions in evolutionary biology is why organisms reproduce sexually. Diploid sexual organisms produce offspring that share only half of their genes with a given parent, and for organisms with two sexes, only the female offspring of sexual organisms with two sexes are directly capable of producing additional offspring. An asexual organism experiences neither of these costs, so the abundance of sexual reproduction in the face of these enormous reproductive costs suggests that in the long term sex must be highly beneficial [reviewed in 72]. Many theories have been put forward to explain the maintenance of sex in the face of such costs [73], and empirical studies of a variety of organisms in the laboratory and in nature have tested these theories [e.g. 74, 75, 76, 77].

One organism that is especially well suited for the study of the costs and benefits of sex is the freshwater New Zealand snail *Potamopyrgus antipodarum*. *P. antipodarum* features both obligately sexual, diploid populations and obligately asexual, polyploid lineages. Sexual and asexual snails are often found in the same lake under equivalent ecological conditions, and there appears to have been many independent derivations of asexual lineages from sexual ancestors [78]. These factors make *P. antipodarum* unique among species used as models for investigating the costs and benefits of sexual versus asexual reproduction.

In order to understand these costs and benefits in *P. antipodarum*, it is necessary to have a clear picture of the evolutionary history of reproductive mode in the species. A previous analysis of mitochondrial gene sequences suggests that asexual lineages are derived from sexual ancestors and that these transitions have occurred several times, independently in different lakes [78]. This same study found that the timing of these transitions was highly variable between lineages, with some lineages apparently derived from sexual ancestors in the ancient past. This is notable because it is often assumed that asexual lineages rapidly accumulate deleterious mutations and are thus evolutionary "dead ends" [79]. These conclusions were based on mitochondrial clock calculations, so they give only a rough account of the history asexual reproduction in the species.

In this chapter, I present a method for inferring the time of transition from sexual to asexual reproduction in triploid organisms, jointly with the population size history of the diploid sexual ancestor of the asexual lineage. The method uses a hidden Markov model (HMM) based on the SMC' model of recombination and coalescence [11], and it takes as input the unphased genome sequence of a triploid organism. An effort to sequence the complete genomes of more than 20 sexual and asexual *P. antipodarum* lineages is currently underway, including 12 triploid asexual lineages. The first public assembly is expected within two months, and I am a part of the team sequencing the genomes and have access to the sequence data. Below, I present the method, which I call "triploid SMC" or "TSMC", and then demonstrate that it is able to recover the transition time and sexual population history in simulated data. I also discuss the possibility of adding additionally biologically relevant features to the inference model, including an initial period of diploid asexuality before a change in ploidy to triploid asexuality and the action of gene conversion in the asexual lineage.

This work is ongoing and at this point only theoretical. When the genome sequences become available, I will apply the TSMC method to the triploid asexual lineages to infer the range of transition times in *P. antipodarum*. This will be a part of a more expansive manuscript on the demographic and phylogeographic history of *P. antipodarum*, to be written in collaboration with Peter Fields. As we are presenting this work prior to publication in a peer-reviewed journal, we request that anyone wishing to use the ideas in this chapter first contact the authors.

4.1 Theory and Results

Like a number of other recent demographic inference methods [e.g., 12, 13, 17], we infer the demographic history of a population by constructing a hidden Markov model (HMM) in which the hidden variable at each point along the genome is the local gene genealogy describing the ancestry of the sampled genomes at that position, and the observed variable is the alignment of the sampled genomes at that point. In our case, we are modeling the genome of a triploid organism, so each hidden state represents a gene genealogy with three leaves.

Define $N(t)$ as the size of the (diploid) ancestral sexual population at time t, going back in time, with t in units of $2N(0)$. Let $N(0)$ be the size of this population at the point in time when the focal triploid asexual lineage was formed from three chromosomes sampled randomly from the diploid, sexual population. Let $\lambda(t) = N(t)/N(0)$ be the relative population size at time t.

At each location along the genome, the gene genealogy will look like the example genealogy shown in Figure 4.1. There will be some period of length T_d, measured in units of $2N(0)$, between the present and the time at which the triploid clonal lineage was formed. During this interval, the three branches of the genealogy are frozen together in the same clonal lineage, undergoing no coalescence and no recombination. (This ignores the possibility of gene conversion in the asexual lineage, which is discussed below.) Going further back in time past the sexual-to-asexual transition time, in the sexual ancestral population, each pair of lineages will coalesce at rate $\lambda(t)^{-1}$, as in a typical coalescent model.

The triploid genomes will be unphased, so we need to incorporate an averaging across phasing into our model. This is accomplished by recording only the first and second coalescence time (in the three-leaved genealogy) as the hidden state at each position. These two coalescence times, with the triploid divergence time T_d (which is the same everywhere along the genome), are sufficient to calculate the emission probabilities of observing a particular alignment of the sequences at a position in the genome.

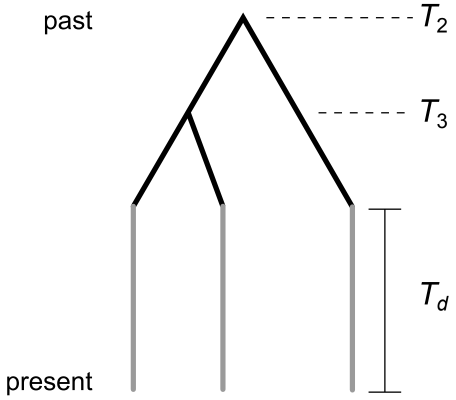

Figure 4.1: Example gene genealogy at a focal point along a triploid asexual genome. During the asexual phase of the lineage's history, no coalescence or recombination occurs and the three sampled lineages are captive together in the same lineage of clonal individuals (gray lines). Going back in time, after the transition between sexual and asexual reproduction, coalescence and recombination occurs according to standard coalescent models.

Define $\Omega(u, v)$ as the cumulative coalescent rate between times u and v:

$$\Omega(u, v) = \int_u^v \frac{dt}{\lambda(t)}. \tag{4.1}$$

The state of the TSMC at each point along the genome is described by the vector $\boldsymbol{t} = (t_3, t_2)$, where t_3 is the time of the first coalescence event and t_2 is the time of the second coalescence event amongst the three lineages in a triploid genome, each measured from the sexual to asexual transition time. The equilibrium joint distribution of (t_3, t_2) is

$$\pi(t_3, t_2) = \frac{3}{\lambda(t_3)\lambda(t_2)} e^{-3\Omega(0,t_3)} e^{-\Omega(t_3,t_2)}. \tag{4.2}$$

We use the SMC' [11] to calculate the transition kernel between different local genealogies (pairs of coalescence times) across the genome. We average across phasing by considering all the different ways that a transition under the SMC' could change the vector of coalescence times (s_3, s_2) to (t_3, t_2) at a recombination site. Let $q(\boldsymbol{t}|\boldsymbol{s})$ be this transition kernel. The following gives the transition densities in terms of unsolved integrals. In each expression, integration is performed over possible locations of recombination events that could lead to the indicated change in states.

$q(\boldsymbol{t}|\boldsymbol{s}) =$

For $t_3 = s_3; t_2 > s_2$:

$$\int_0^{s_3} \frac{du}{2s_2 + s_3} e^{-3\Omega(u,s_3)} e^{-2\Omega(s_3,s_2)} \frac{1}{\lambda(t_2)} e^{-\Omega(s_2,t_2)} + 2\int_{s_3}^{s_2} \frac{du}{2s_2 + s_3} e^{-2\Omega(u,s_2)} \frac{1}{\lambda(t_2)} e^{-\Omega(s_2,t_2)}$$

For $t_3 = s_3; t_2 < s_2$:

$$\int_0^{s_3} \frac{du}{2s_2 + s_3} e^{-3\Omega(u,s_3)} \frac{1}{\lambda(t_2)} e^{-2\Omega(s_3,t_2)} + 2\int_{s_3}^{t_2} \frac{du}{2s_2 + s_3} \frac{1}{\lambda(t_2)} e^{-2\Omega(u,t_2)}$$

For $t_3 < s_3; t_2 = s_3$:

$$\int_0^{t_3} \frac{du}{2s_2 + s_3} e^{-3\Omega(u,t_3)} \frac{2}{\lambda(t_3)}$$

For $t_3 < s_3; t_2 = s_2$:

$$2\int_0^{t_3} \frac{du}{2s_2 + s_3} \frac{2}{\lambda(t_3)} e^{-3\Omega(u,t_3)}$$

For $t_3 > s_3; t_2 = s_2$:

$$2\int_0^{s_3} \frac{du}{2s_2 + s_3} e^{-3\Omega(u,s_3)} \frac{2}{\lambda(t_3)} e^{-2\Omega(s_3,t_3)}$$

For $t_3 = s_2; t_2 > s_2$:

$$2\int_0^{s_3} \frac{du}{2s_2+s_3} e^{-3\Omega(u,s_3)} e^{-2\Omega(s_3,s_2)} \frac{1}{\lambda(t_2)} e^{-\Omega(s_2,t_2)}$$

For $t_3 = s_3; t_2 = s_2$:

$$\begin{aligned}3\int_0^{s_3} \frac{du}{2s_2+s_3}\frac{1}{3}\left[1-e^{-3\Omega(u,s_3)}\right] + \int_0^{s_3}\frac{du}{2s_2+s_3} e^{-3\Omega(u,s_3)}\frac{1}{2}\left[1-e^{-2\Omega(s_3,s_2)}\right] \\ + 2\int_{s_3}^{s_2}\frac{du}{2s_2+s_3}\frac{1}{2}\left[1-e^{-2\Omega(u,s_2)}\right].\end{aligned} \tag{4.3}$$

Each part is implicitly multiplied by a delta function to limit the density to points where the parameters are assumed to be equal to each other. For example, the first part of $q(\boldsymbol{t}|\boldsymbol{s})$ is implicitly multiplied by $\delta(t_3 - s_3)$, and the last part is multiplied by $\delta(t_3 - s_3)\delta(t_2 - s_2)$.

4.2 Piecewise constant transition probabilities

Like other HMM methods based on the sequentially Markov coalescent [12, 13, 17, but see also 80], we will assume that the size of the sexual, ancestral population is piecewise constant, such that the population changes size at times $(T_1, \ldots, T_n)$ and the size between T_i and T_{i+1} is a constant $N(0)\lambda_i$. Define $T_0 = 0$, $T_{n+1} = \infty$ and $\Delta_i = T_{i+1} - T_i$. Let $\alpha(t)$ be the index of the time interval to which t belongs, *i.e.*, $\alpha(t) = \max_i\{i : T_i \leq t\}$. Then the cumulative coalescent rate between u and v can be written

$$\Omega(u,v) = \begin{cases} \frac{v-u}{\lambda_{\alpha(u)}} & \alpha(u) = \alpha(v) \\ \\ \frac{T_{\alpha(u)+1}-u}{\lambda_{\alpha(u)}} + \sum_{i=\alpha(u)+1}^{\alpha(v)-1} \frac{\Delta_i}{\lambda_i} + \frac{v-T_{\alpha(v)}}{\lambda_{\alpha(v)}} & \alpha(u) < \alpha(v). \end{cases} \tag{4.4}$$

Under a piecewise constant population size history, the equilibrium joint density of (t_3, t_2) is now approximately

$$\pi(t_3, t_2) = \frac{3}{\lambda_{\alpha(t_3)}\lambda_{\alpha(t_2)}} e^{-3\Omega(0,t_3)} e^{-\Omega(t_3,t_2)}. \tag{4.5}$$

Assuming a piecewise constant population size history allows the integrals in (4.3) to be written in terms of simple functions. We present these equations in Appendix F.

There are several integrals of the form $\int_x^y e^{-k\Omega(u,y)}du$ in Equation (4.3). Assuming a piecewise population size history, this integral can be written

$$\int_x^y e^{-k\Omega(u,y)}du = \\ e^{-k\Omega(T_{\alpha(x)+1},y)} \left[1 - e^{-\frac{k\left(T_{\alpha(x)+1}-x\right)}{\lambda_{\alpha(x)}}}\right] \frac{\lambda_{\alpha(x)}}{k} + \sum_{i=\alpha(x)+1}^{\alpha(y)-1} e^{-k\Omega(T_{i+1},y)} \left[1 - e^{-\frac{k\Delta_i}{\lambda_i}}\right] \frac{\lambda_i}{k} + \\ \left[1 - e^{-\frac{k\left(y-T_{\alpha(y)}\right)}{\lambda_{\alpha(y)}}}\right] \frac{\lambda_{\alpha(y)}}{k}. \tag{4.6}$$

The full derivation of this equation is given in Appendix E.

With this equation, we can calculate all of the transition probabilities in the transition kernel (4.3). These are given in Appendix G.

4.3 Discrete approximation to the triploid SMC' coalescent process

In order to construct a hidden Markov model from these transition densities, it is necessary to discretize the coalescent process described above. Let the discrete state (i, j), $i \leq j$, correspond to the continuous states in which $T_i < t_3 < T_{i+1}$ and $T_j < t_2 < T_{j+1}$. We first calculate the equilibrium probability $\pi_{i,j}$ that the continuous-time coalescent process with piecewise population history is in $(i, j), i < j$:

$$\pi_{i,j} = \frac{3}{2} e^{-3\Omega(0,T_i)} e^{-\Omega(T_{i+1},T_j)} \left(e^{\frac{-\Delta_i}{\lambda_i}} - e^{\frac{-3\Delta_i}{\lambda_i}} \right) \left[1 - e^{-\frac{\Delta_j}{\lambda_j}} \right]. \tag{4.7}$$

If $j = n$, we let $\Delta_j = \infty$ and $1 - \exp(-\Delta_j/\lambda_j) = 1$. We also calculate $\pi_{i,i}$:

$$\pi_{i,i} = \frac{1}{2} e^{-3\Omega(0,T_i)} \left(2 - 3e^{-\frac{\Delta_i}{\lambda_i}} + e^{-\frac{3\Delta_i}{\lambda_i}} \right). \tag{4.8}$$

For $i = n$, again we let $\Delta_i = \infty$ and thus $\pi_{n,n} = \exp(-3\Omega(0, T_n))$.

To calculate the transition probabilities at recombination sites under our discrete approximation to the continuous-time coalescent process, we will assume that when the process is in the state (i, j) (as described above), the local coalescence times are $(\mathrm{E}_{i,j}[t_3], \mathrm{E}_{i,j}[t_2])$, where $\mathrm{E}_{i,j}[t_3]$ and $\mathrm{E}_{i,j}[t_2]$ are the marginal expected coalescence times of t_3 and t_2, respectively, under the continuous-time, piecewise-constant population history model. The marginal expectation of t_3 in the interval $(i, j), i < j$ is

$$\mathrm{E}_{i,j}[t_3] = \frac{3}{4\pi_{i,j}} e^{-3\Omega(0,T_i)} \left(1 - e^{-\frac{\Delta_j}{\lambda_j}} \right) e^{-\Omega(T_{i+1},T_j)} \left[(\lambda_i + 2T_i) e^{\frac{-\Delta_i}{\lambda_i}} - (\lambda_i + 2T_{i+1}) e^{\frac{-3\Delta_i}{\lambda_i}} \right]. \tag{4.9}$$

With $j = n$, this is

$$\mathrm{E}_{i,j}[t_3] = \frac{3}{4\pi_{i,n}} e^{-3\Omega(0,T_i)} e^{-\Omega(T_{i+1},T_n)} \left[(\lambda_i + 2T_i) e^{\frac{-\Delta_i}{\lambda_i}} - (\lambda_i + 2T_{i+1}) e^{\frac{-3\Delta_i}{\lambda_i}} \right]. \tag{4.10}$$

The marginal expectation of t_2 in $(i, j), i < j$ is

$$\mathrm{E}_{i,j}[t_2] = \frac{3}{2\pi_{i,j}} e^{-3\Omega(0,T_i)} e^{-\Omega(T_{i+1},T_j)} \left(e^{\frac{-\Delta_i}{\lambda_i}} - e^{\frac{-3\Delta_i}{\lambda_i}} \right) \left(\lambda_j + T_j - (\lambda_j + T_{j+1}) e^{-\frac{\Delta_j}{\lambda_j}} \right), \tag{4.11}$$

and with $j = n$, this is

$$\mathrm{E}_{i,n}[t_2] = \frac{3}{2\pi_{i,n}} e^{-3\Omega(0,T_i)} e^{-\Omega(T_{i+1},T_n)} \left(e^{\frac{-\Delta_i}{\lambda_i}} - e^{\frac{-3\Delta_i}{\lambda_i}} \right) (\lambda_n + T_n) \tag{4.12}$$

We must also calculate the marginal expectations of t_3 and t_2 in the state (i, i):

$$\mathrm{E}_{i,i}[t_3] = \frac{1}{12\pi_{i,i}} e^{-3\Omega(0,T_i)} \left(4(3T_i + \lambda_i) + e^{-\frac{3\Delta_i}{\lambda_i}} (6T_{i+1} + 5\lambda_i) - 9e^{-\frac{\Delta_i}{\lambda_i}} (2T_i + \lambda_i), \right) \tag{4.13}$$

and for $i = n$, this expectation is

$$\mathrm{E}_{n,n}[s_3] = T_n + \frac{\lambda_n}{3}. \tag{4.14}$$

The marginal expectation of t_2 in the discrete state $(i, i), i < n$ is

$$\mathrm{E}_{i,i}[t_2] = \frac{1}{6\pi_{i,i}} e^{-3\Omega(0,T_i)} \left(e^{-\frac{3\Delta_i}{\lambda_i}} (3T_{i+1} + \lambda_i) + 6T_i + 8\lambda_i - 9e^{-\frac{\Delta_i}{\lambda_i}} (T_{i+1} + \lambda_i) \right), \tag{4.15}$$

and for $i = n$, this expectation is

$$\mathrm{E}_{n,n}[s_2] = T_n + \frac{\lambda_n}{3} + \lambda_n. \tag{4.16}$$

4.3.1 Discrete $q\big((k,l)|(i,j)\big)$ transition function

To calculate the transition probabilities from (discrete-time) state (i,j), to state (k,l), we integrate the continuous-time transition kernel (see eqn. 4.6) over the interval corresponding to (k,l), replacing s_3 and s_2 with their conditional expectations $\mathrm{E}_{i,j}[s_3]$ and $\mathrm{E}_{i,j}[s_2]$ respectively. Thus

$$q\Big((k,l)\,|\,(i,j)\Big) = \int_{T_k}^{T_{k+1}} \int_{T_l}^{T_{l+1}} q\Big((t_3,t_2)|\,(\mathrm{E}_{i,j}[s_3], \mathrm{E}_{i,j}[s_2])\,\Big)dt_2 dt_3. \tag{4.17}$$

Note that in any single transition, either the first or second coalescence time changes, but not both. This simplifies the calculation of these integrals. The discrete-time transition probabilities are given in Appendix G.

4.3.2 Emission probabilities

We encode the genotype at every position in the genome as one of three different values: 0, 1, and 2. The state 0 represents a homozygous site, and 1 (2) represent sites where one (two) of the three chromosomes have a derived (*i.e.*, non-ancestral) copy at that position.

To form the observed chain in our HMM, we consider all the genotypes in a stretch of b base pairs and categorize that stretch of the sequence with a state 0, 1, 2, or 3. The state 0 means that the stretch of b base pairs is completely homozygous. The state 1 means that there is at least one site that has a 1 genotype and none that have a 2 genotype. Likewise, the state 2 means that at least one site has a 2 genotype, and none have a 1 genotype. The state 3 means that at least one site had a 1 genotype and at least one site had a 2 genotype.

With observed states coded this way, the emission probabilities given local coalescence times t_3 and t_2 are

$$e_k(t_3, t_2, T_d) = \begin{cases} e^{-\frac{\theta b\left(2t_2+t_3+3T_d\right)}{2}} & k=0 \\ e^{-\frac{\theta b(t_2-t_3)}{2}}\left(1-e^{-\frac{\theta b\left(2t_3+t_2+3T_d\right)}{2}}\right) & k=1 \\ \left(1-e^{-\frac{\theta b(t_2-t_3)}{2}}\right)e^{-\frac{\theta b\left(2t_3+t_2+3T_d\right)}{2}} & k=2 \\ \left(1-e^{-\frac{\theta b(t_2-t_3)}{2}}\right)\left(1-e^{-\frac{\theta b\left(2t_3+t_2+3T_d\right)}{2}}\right) & k=3. \end{cases} \tag{4.18}$$

As above, T_d is the asexual divergence time, *i.e.*, the time in the past when the asexual lineage was derived from a sexual ancestor. The above probabilities assume that t_3 and t_2 are measured continuously. In practice, we discretize time, so for a particular hidden state (i, j), we replace t_3 with $\mathrm{E}_{i,j}[t_3]$ and t_2 with $\mathrm{E}_{i,j}[t_2]$.

Classifying observed states this way requires that each polymorphism be polarized against an outgroup. If this is not possible, then the states can be recoded as 0 and 1, where 0 is a stretch of b completely homozygous base pairs, and 1 is a stretch of b base pairs with at least one polymorphic position. In this case the probabilities become

$$e_k(t_3, t_2, T_d) = \begin{cases} e^{-\frac{\theta b\left(2t_2+t_3+3T_d\right)}{2}} & k=0 \\ 1-e^{-\frac{\theta b\left(2t_2+t_3+3T_d\right)}{2}} & k=1. \end{cases} \tag{4.19}$$

The parameter b can be tuned to match the observed polymorphism. If the change in ploidy T_d generations ago also involved a change in mutation rate, this new mutation rate will be unidentifiable, that is, impossible to distinguish from a proportionally scaled T_d. Thus T_d should be viewed as a compound parameter.

4.4 Description of Hidden Markov Model

The transition probabilities $q\big((k,l)|(i,j)\big)$ are the probabilities of transitioning from state (i,j) to state (k,l) at sites of ancestral recombination. They can be used to obtain the the transition matrix $\big\{P_{(i,j),(k,l)}\big\}$, where

$$P_{(i,j),(k,l)} = \left(1 - e^{-\frac{\rho}{2}\left(2\,\mathrm{E}_{i,j}[t_2]+\mathrm{E}_{i,j}[t_3]\right)}\right) q\big((k,l)|(i,j)\big) \tag{4.20}$$

is the probability of transitioning from state (i,j) to state (k,l), $(i,j) \neq (k,l)$, unconditional on there being an ancestral recombination event at this site. This follows from the fact that recombination events are encountered across the genome at rate $\frac{\rho}{2}\,(2\,\mathrm{E}_{i,j}[t_2] + \mathrm{E}_{i,j}[t_3])$ when the local state is (i,j). Diagonal entries of this matrix are obtained by subtracting the sum of the off-diagonal elements from 1.

We define a hidden Markov chain $\{X_i\}$ that is governed by this transition matrix. The observed process $\{Y_i\}$ represents the emissions, taking values 0 through 3 as described above. We use the EM algorithm to iteratively maximize the expectation of the full likelihood

$$P(X,Y|\theta) = e_{x_1}(y_1)\,\pi_{x_1} \prod_{i=1}^{T-1} P_{x_i,x_{i+1}} e_{x_{i+1}}(y_{i+1}), \tag{4.21}$$

where y_i is the observed state at position i, and x_i is the state of the hidden chain at position i, and T is the length of the sequence. In practice we maximize the log-likelihood

$$\log P(X,Y|\theta) = \log\left(e_{x_1}(y_1)\right) + \log(\pi_{x_1}) + \sum_{i=1}^{T-1} \log\left(P_{x_i,x_{i+1}}\right) + \log\left(e_{x_{i+1}}(y_{i+1})\right). \tag{4.22}$$

We integrate over the states of the hidden chain by pairing the expectation-maximization (EM) algorithm with the forward-backward algorithm for calculating likelihoods with these chains, modified to avoid underflow errors. Following Li and Durbin [12], we constrain the number of free population size parameters to be less than the number of discretized time intervals, such that population sizes are repeated over multiple adjacent time intervals according to a user-specified pattern. As in [12], the boundaries of the time intervals were placed at

$$t_i = 0.1(e^{\frac{i}{n}\log(1+10T_{max})} - 1) \tag{4.23}$$

for intervals $i \in \{0, 1, \ldots, n\}$. We tested additional spacing schemes for these time intervals and found that they had little effect on the accuracy of inference.

In summary, the free parameters inferred by the triploid SMC inference procedure include: the recombination rate ρ, scaled by $4N(0)$ and implicitly multiplied by b, the number of base pairs considered in a single emission state; the mutation rate θ, scaled by $4N(0)$; the triploid transition time T_d, implicitly proportional to any change in the genomic mutation rate since the transition from sexual to asexual reproduction; $\{\lambda_i\}$, the free population size parameters, constrained such that the first, λ_0, is equal to 1; and T_{max}, the lower boundary of the final time interval (as used in Eq. 4.23), included to improve the fit of the inferred population sizes at discretized intervals to the actual population history.

Because the number of states (i, j) grows quadratically in the number of discretized time intervals n, the complexity of the forward-backward algorithm for the TSMC is $O(n^4L/b)$, where L is the number of base pairs considered. We find that the TSMC requires substantially greater runtimes than the PSMC, for which the forward-backward algorithm is of complexity $O(n^2L/b)$. In practice, this requires us to limit the number of discretized time intervals. It may be possible to reduce the complexity of the forward-backward algorithm by exploiting symmetries in the coalescent process, as demonstrated by [81], but we find that the running times and accuracy of inference are reasonable without these optimizations.

4.4.1 Inference on Simulated Triploid Asexual Genomes

We tested the triploid SMC inference procedure using simulated triploid genomes. To generate triploid genomes, we simulated three chromosomes under the standard neutral, sexual coalescent with recombination [30] and then added singleton mutations (accumulated after the transition to asexual reproduction, sometimes referred to as the "Meselson effect" [e.g., 82]) at Poisson rate $3\theta T_d/2$ uniformly across the genome and across the three chromosomes to simulate the effects of the lineages being captive together in the same asexual lineage for a time interval of length T_d. Simulations were carried out using MSPRIME [83] and custom scripts.

We simulated genomes of asexual triploid lineages of different ages and having sexual ancestors with different population size histories. Figure 4.2 shows the inferred sexual-to-asexual transition time and ancestral population sizes for four different asexual lineage ages (0, 5000, 10000, and 20000 generations), a sexual ancestral population size of $N = 10000$, a mutation rate of 1.5×10^{-8} per base pair per generation, and a total genome length of 100 Mbp. The EM algorithm was run for 20 iterations and three free population sizes were inferred. Time was discretized into 16 intervals. The inferred transition times and population sizes were fairly accurate, but the procedure showed a slight bias towards underestimating the age of the asexual lineage and overestimating the sexual population size at the onset of asexual reproduction.

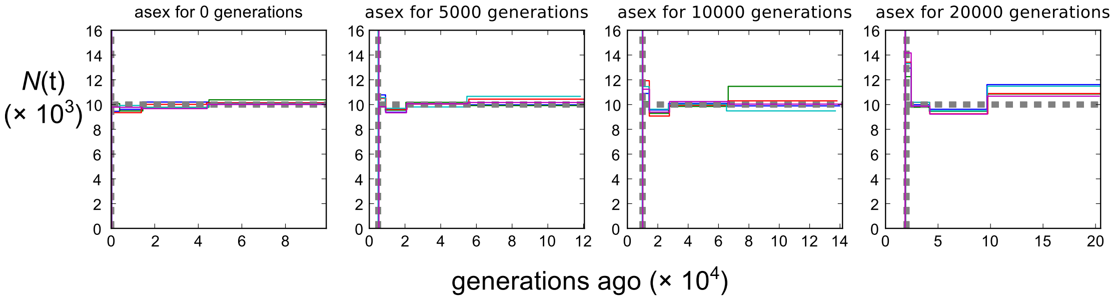

Figure 4.2: Inferred demographic history of asexual lineages of different ages with constant-sized ancestral sexual populations. Each panel shows the inferred sexual-to-asexual transition time and population size history of the sexual ancestor for five replicate simulations. The true values are shown with a thick gray dashed line, and the inferred history of each replicate simulation is shown with a differently colored thin solid line. Vertical lines show the inferred transition time, and horizontal lines to the right of the transition show the inferred population size history of the sexual ancestor. In each simulation, the size of the sexual ancestral population was $N = 10000$, the mutation rate was 1.5×10^{-8} per generation per base pair, and the recombination rate was 1.0×10^{-8} per generation per base pair. Each simulated genome was 120 Mbp in length.

In order to test whether inference of the sexual-to-asexual transition time is confounded by variable population sizes in the ancestral sexual population, we simulated scenarios in which the sexual ancestral population changed in size. When we simulated the demographic history given as an example in Figure 2 of [12], featuring three changes in population size and including a severe bottleneck, the timing and magnitude of these transitions were well inferred, on average, as was the timing of the transition to asexual reproduction (Fig. 4.3A). On the other hand, in simulations in which the asexual lineage was derived from a sexual population undergoing extreme exponential growth, the TSMC showed a slight bias towards overestimating the age of the asexual lineages (Fig. 4.3B). This can be explained by the fact that coalescence is infrequent in very large populations, just as it does not occur at all in asexual lineages. There is presumably some information in linkage patterns that can be used to distinguish an exponentially growing sexual ancestral population from an extension of the asexual phase of reproduction, since recombination does occur in the sexual ancestor, no matter the size, and

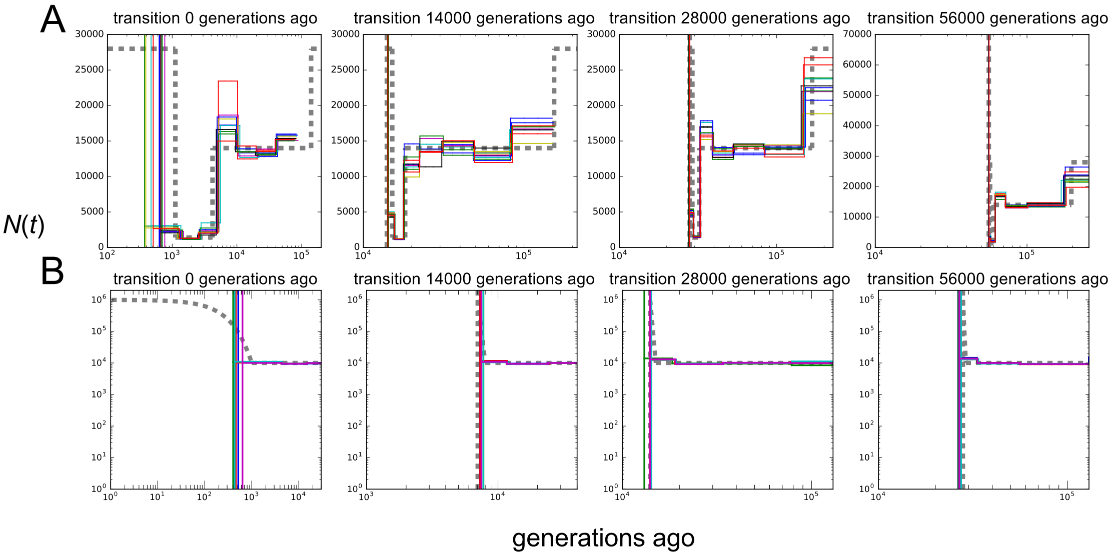

Figure 4.3: Inferred demographic history of asexual lineages whose sexual ancestors experienced a population bottleneck or exponential growth before the onset of asexual reproduction. True (gray lines) and inferred (colored lines) population sizes are displayed as in Figure 4.2. In each simulation, the mutation rate is 1.5×10^{-8} per base pair per generation, and the recombination rate is 1.0×10^{-8}, with a total genome size of 120 Mbp. The top row **(A)** shows inferred population histories for asexual lineages whose ancestors underwent a population bottleneck prior to the onset of asexual reproduction. The bottom row **(B)** shows inferred population histories for asexual lineages whose sexual ancestral population was growing exponentially at the onset of asexual reproduction.

does not occur in the asexual lineage. However, it seems that the TSMC cannot use this information to distinguish asexual reproduction from a rapidly growing sexual ancestral population.

We tested the power to distinguish asexual lineage ages on a finer scale by carrying out simulations where the asexual lineages varied between 0 and 1000 generations in age, with a sexual ancestral population of constant size $N = 10000$. The TSMC tended to overestimate the age of the asexual lineage by a few hundred generations (Fig. S7). In these simulations, at most 5% of the heterozygous sites, on average, are due to mutations during the asexual period, and there is little signal to distinguish lineages of different ages. In simulations where the the asexual lineage is young and the sexual ancestor underwent a bottleneck prior to the onset of the asexual lineage, the TSMC was unable to distinguish between the large recent population size in the sexual ancestor and an extension of the period of asexual reproduction (Fig. S8). Thus care is needed in the interpretation of the inferred lineage age for any recently derived triploid asexual lineage.

In simulations, we know the true per-generation mutation rate, so we can scale by the actual value to write scaled parameters in units of generations. In a real application to data from *P. antipodarum*, the true mutation rate is unknown, so inferred parameters will be scaled by an unknown population size. Thus it is most realistic to view our application of the TSMC as a way of generating a qualitative understanding of the different timings of transitions from sexual to asexual reproduction in triploid asexual *P. antipodarum* lineages.

4.4.2 Diploid asexual lineages and transitions from diploid to triploid asexual reproduction

Across a range of species, asexual reproduction is associated with polyploidy, and in *P. antipodarum* most of the known asexual lineages are triploid or tetraploid. However, there is one diploid isofemale lineage that has never produced a male in the laboratory and is thus putatively asexual (M. Neiman, personal communication). Thus one hypothesis is that asexual lineages tend to form as diploids and then rapidly transition from diploidy to triploidy as asexual lineages after being triploidized by a sexual haploid sperm. A subsequent transition to tetraploidy may be caused by the incorporation of an additional haploid sperm into the genome.

Conveniently, the age of a diploid asexual lineage can be inferred using the standard PSMC method [12], which features an option that is perfectly suited for inferring the age of a diploid asexual lineage. This is the `-T` option (called the "divergence time" parameter by the authors), which is disabled by default and appears to have been intended to infer the timing of divergence between two isolated populations. If this is the intended usage, it does not allow for recombination during the period of isolation (as it should) and will produce unreliable estimates in that context. However, it is perfectly suited for studying the age of diploid asexual lineages, and accurately recovers the age of a diploid asexual lineage in simulations (not shown). We will apply the PSMC with this option enabled to the diploid genomes from the *P. antipodarum* genome sequencing project.

To test the hypothesis that triploid asexual lineages are derived from diploid asexual lineages via incorporation of haploid sperm from a sexual male, we modified the transition probabilities of our hidden Markov chain to reflect this scenario. In particular, we allowed for an additional period in which two chromosomes are frozen together in the same (diploid) asexual lineage and a third chromosome is found in a sympatric sexual population, undergoing recombination (Fig. 4.4). The updated transition probabilities are given in Appendix H. In the version of the TSMC with this additional diploid asexual period, in addition to inferring the total asexual time T_d, we also infer the time of the transition to triploid asexual reproduction, D_3. For notational convenience, we assume that the first transition, to diploid asexual reproduction, happens at time $D_2 = 0$ and then the subsequent transition to triploid asexual reproduction happens at time $D_3 < 0$ so that $-D_3$ is actually the length of the diploid asexual interval (see 4.4). This allows us to measure time in the diploid sexual population starting from $t = D_2 = 0$.

Testing the modified TSMC with simulated genomes of triploid asexual lineages that were first diploid asexual lineages, we found that the the timing of the two transitions were accurately recovered only when the total asexual time was small and the triploid

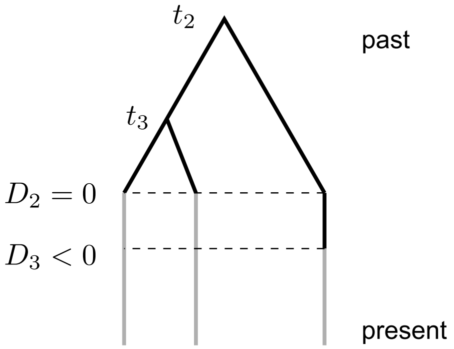

Figure 4.4: Example gene genealogy for an asexual lineage that was diploid and later become triploid by incorporation of a haploid, sexual sperm. T_3 and T_2 represent the two coalescence times in the interval in which all three chromosomal lineages are sexual, as before. The time $D_2 = 0$ is the time of the initial transition to sexual reproduction, and $D_3 < 0$ is the (negative of the) length of the interval before the diploid asexual lineage transitions to triploid asexual reproduction, measured in units of $2N(0)$, the sexual population size at the time of transition to diploid asexual reproduction from sexual reproduction. In the figure, black lines show the sexual branches of the gene genealogy that contribute to the total rate of recombination.

period was substantially longer than the diploid period (Fig. S9). In other scenarios, the relative timing of the diploid and triploid transitions was not reliably inferred. These changes to the SMC' calculations underlying the HMM transition probabilities affect only the linkage patterns in the model — the marginal distribution of gene genealogies remains the same. Furthermore, the additional recombination allowed during the period of diploid asexual reproduction makes only minor changes to the linkage patterns. Given that SMC-based methods are known to be poor at inferring recombination rates [12, 17, also observed with TSMC], it is perhaps not surprising that the TSMC does not do well at inferring these additional parameters. However, it is encouraging that the *total* time spent as an asexual lineage is still inferred accurately with the inclusion of separate diploid and triploid asexual transition times (Fig. S9).

4.5 Discussion

We have created a SMC-based HMM model for demographic inference with triploid genomes. To our knowledge, it is the first SMC-based HMM demographic inference procedure that models the full gene genealogy under the standard coalescent for more than two sequences. The Bayesian MCMC method of Palacios et al. [80] models the genealogies of many chromosomes under the SMC', but it takes an ancestral recombination graph as input rather than genetic sequence data. DiCal [13] models the genealogies of multiple chromosome sequences, but it bases its calculations on a genealogical process approximating the coalescent, derived from an approximation to the standard Wright-Fisher diffusion. MSMC [17] allows multiple sequences but models only the most recent coalescence time in its HMM. We are able to model the full coalescent with three sequences in part because we average over phasing. This permits us to model only the first and second coalescence time in a genealogy with three leaves, reducing the number of states in our HMM by a factor of six. Surely much information is lost by discarding phasing, but since the triploid *P. antipodarum* genomes will be unphased, in our application there would be no point to modeling phased genomes.

We have neglected to account for the possibility of gene conversion in the asexual *P. antipodarum* lineages. In the context of sexually reproducing organisms, gene conversion can be modeled as a special type of recombination [84]. In the context of clonal asexual lineages, gene conversion occurring in an asexual lineage acts analogously to coalescence, since each gene conversion event copies one sampled, captive chromosome onto another. Flot et al. [85] studied this process in (diploid) bdelloid rotifers and used Monte Carlo simulations to infer parameters related to gene conversion. Recently, Hartfield et al. [86] studied theoretical single-locus coalescence patterns in facultatively reproducing organisms undergoing gene conversion. In order to incorporate gene conversion into an SMC-based HMM model like the PSMC or TSMC, it would be necessary to have a sequentially Markov model of gene conversion in clonally reproducing asexual organisms. The gene conversion process is not Markov [87], but a first-order Markov approximation can be made. In diploid organisms, it is necessary to model only the local pairwise coalescence time, as in the PSMC [12], and transition probabilities and two-locus properties should be relatively straightforward to calculate after assuming that the process is first-order Markov. To include gene conversion in the TSMC model, it would be necessary to create a triploid version of the gene conversion ancestry model, which is substantially more complex than the diploid version.

Although it is likely that asexual *P. antipodarum* experience some gene conversion, and it would be of biological interest to infer the parameters of gene conversion in this species, at present we do not to model gene conversion in the TSMC. We predict that any inferred asexual lineage ages will be biased downwards in the presence of gene conversion, since any gene conversion that occurs during the asexual phase will look like coalescence, which we assume can occur only during the sexual phase of reproduction in the TSMC. Ancient asexual lineages have been proposed in *P. antipodarum* [78], and it is ancient asexual lineages that are supposed to be unlikely under the assumption that asexual reproduction is an evolutionary dead end [79]. Since gene conversion will likely cause a downward bias in the estimated age of an asexual lineage, we view the TSMC as being conservative in identifying ancient asexual lineages in the presence of gene conversion in asexual lineages.

We expect that the TSMC will be useful in the study of other organisms featuring triploid apomictic asexual reproduction. Triploid apomictic reproduction is encountered in a variety of taxa, including lizards [88], fish [89], and various plants [e.g. 90, 91, 92].

Many other apomictic asexual organisms are diploid or tetraploid. To infer the history of asexual reproduction in diploid asexual lineages, it is possible to use the PSMC with the `-T` option enabled, as discussed above. (Additional modifications would need to be made to properly accommodate gene conversion.) An SMC-based HMM approach could be devised to study the evolutionary and population genetic history of tetraploid asexual lineages, but modeling the genealogies fully would involve a significant increase in the number of states in the HMM. Without phasing, the number of states to consider would increase by a factor of approximately $2n$ (where n is the number of discrete time intervals) over our implementation of the TSMC, and with phasing it would increase by substantially more. This would increase runtimes substantially, although, again, improvements may be possible by exploiting symmetries in the coalescent [81].

Besides gene conversion, one possibility that has not been addressed here is the occurrence of cryptic or infrequent sexual reproduction. If there is a history of infrequent sex in an asexual lineage, any inference procedure involving a coalescent-based HMM would need to be adapted to allow for the possibility of additional recombination and coalescence during the asexual part of a lineage's history. Exactly how this could be achieved depends on whether the occasional sex is with sexual individuals or other mostly asexual individuals. We expect that either situation could be accommodated by a suitably modified SMC model, but we leave this for future work.

A

Pedigrees and biased estimators of θ

Estimates of the population-scaled mutation rate $\theta = 4N\mu$ will be downwardly biased if there is recent IBD in the sample, since sequences will be identical with an artificially inflated probability, and this resembles coalescence prior to any mutation between the two identical sequences.

Suppose that we sample two copies of a DNA sequence from each of n diploid individuals from a panmictic population. As in the main text of Chapter 1, we index these sequences with $\mathcal{I}_n := \{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}, 2^{\mathrm{p}}, \ldots, n^{\mathrm{m}}, n^{\mathrm{p}}\}$. Let $\mathbb{P}_{\mathcal{I}_n}$ be the set of partitions of $\mathcal{I}_n$. Each pedigree $\mathcal{P}$ induces a set of sample reconfigurations $\mathcal{R}(\mathcal{P}) \subseteq \mathbb{P}_{\mathcal{I}_n}$, where each partition $r \in \mathcal{R}(\mathcal{P})$ represents a possible outcome of segregation through the recent sample pedigree. Each reconfiguration $r \in \mathcal{R}(\mathcal{P})$ contains $|r|$ non-empty, disjoint subsets, each representing a distinct lineage that survives after segregation through the recent pedigree. Associated with each pedigree is also a probability distribution $\Pr(r \mid \mathcal{P}), r \in \mathcal{R}(\mathcal{P})$ representing the Mendelian probabilities of the different sample reconfigurations.

We consider the bias of three estimators of θ that are unbiased in the absence of recent IBD. One estimator of θ we consider is Watterson's [1975] estimator

$$\hat{\theta}_S = \frac{\sum_{i=0}^{L} S^{(i)}}{a_{2n} L}, \tag{A.1}$$

where $2n$ is the (diploid) sample size, $S^{(i)}$ is the number of segregating sites at locus i, L is the number of loci, and $a_{2n} = \sum_{i=1}^{2n-1} 1/i$. The expected value of $\hat{\theta}_S$ given pedigree $\mathcal{P}$ is

$$\begin{aligned} \mathrm{E}\left[\hat{\theta}_S \mid \mathcal{P}\right] &= \sum_{r \in \mathcal{R}(\mathcal{P})} \mathrm{E}\left[\hat{\theta}_S \mid r\right] \Pr(r \mid \mathcal{P}) \\ &= \sum_{r \in \mathcal{R}(\mathcal{P})} \mathrm{E}\left[\frac{S^{(1)}}{a_2 n} \mid r\right] \Pr(r \mid \mathcal{P}) \\ &= \theta \sum_{r \in \mathcal{R}(\mathcal{P})} \frac{a_{|r|}}{a_2 n} \Pr(r \mid \mathcal{R}(\mathcal{P})). \end{aligned} \tag{A.2}$$

This follows from the fact that when there are $|r|$ lineages after segregation through the recent pedigree, the expected number of segregating sites for that sample is $\theta \sum_{i=1}^{|r|-1} \frac{\theta}{i} = \theta a_{|r|}$.

A second estimator of θ is $\hat{\pi}$, the mean number of differences between all pairs of sequences in a sample, which can be written in terms of the site-frequency spectrum:

$$\hat{\pi} = \frac{1}{\binom{2n}{2}} \sum_{i=1}^{2n-1} i(2n-i)\hat{\xi}_i, \tag{A.3}$$

where $\hat{\xi}_i$ is the number of derived alleles present in i out of $2n$ sequences in the sample.

To calculate the expected value of $\hat{\pi}$ given $\mathcal{P}$, it is necessary to consider how IBD in the recent pedigree changes the site frequency spectrum. Define $\mathcal{S}^{(\mathrm{n})} := \{A \subseteq \Omega : \Omega \in \mathbb{P}_{\mathcal{I}_n}\}$ and $\psi : \mathbb{P}_{\mathcal{I}_n} \times \mathbb{N} \to \mathcal{S}^{(\mathrm{n})}$ as

$$\psi(\Omega, i) := \{\omega \subseteq \Omega : \sum_{g \in \omega} |g| = i\}. \tag{A.4}$$

That is, $\psi(\Omega, i)$ is the set of all subsets of the partition Ω such that the total size of all the groups in each subset is i. For example, for $n = 3$ and

$$\Omega = \left\{\{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}\}, \{2^{\mathrm{p}}, 3^{\mathrm{m}}\}, \{3^{\mathrm{p}}\}\right\},$$

we have

$$\begin{aligned}
\psi(\Omega, 1) &= \{\{\{3^{\mathrm{p}}\}\}\} \\
\psi(\Omega, 2) &= \{\{\{2^{\mathrm{p}}, 3^{\mathrm{m}}\}\}\} \\
\psi(\Omega, 3) &= \{\{\{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}\}\}, \{\{2^{\mathrm{p}}, 3^{\mathrm{m}}\}, \{3^{\mathrm{p}}\}\}\} \\
\psi(\Omega, 4) &= \{\{\{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}\}, \{3^{\mathrm{p}}\}\}\} \\
\psi(\Omega, 5) &= \{\{\{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}\}, \{2^{\mathrm{p}}, 3^{\mathrm{m}}\}\}\}.
\end{aligned}$$

Then the expectation of $\hat{\xi}_i$ given reconfiguration $r \in \mathcal{R}(\mathcal{P})$ is

$$\mathrm{E}\left[\hat{\xi}_i \mid r\right] = \sum_{\omega \in \psi(r,i)} \frac{\theta}{|\omega| \binom{|r|}{|\omega|}}, \tag{A.5}$$

since each segregating site that is present in i present-day lineages must have occurred on the branch ancestral to the post-IBD lineages represented by some $\omega \in \psi(r,i)$. The expected number of mutations occurring on a branch subtending $|\omega|$ lineages is $\theta/|\omega|$, and the expected fraction of such mutations that occur on the branch ancestral to the lineages in ω is $1/\binom{|r|}{|\omega|}$, by exchangeability. This gives the expectation of $\hat{\pi}$ conditional on the pedigree:

$$\begin{aligned} \mathrm{E}\left[\hat{\pi} \mid \mathcal{P}\right] &= \sum_{r \in \mathcal{R}(\mathcal{P})} \mathrm{E}\left[\hat{\pi} \mid r\right] \Pr(r \mid \mathcal{P}) \\ &= \frac{1}{\binom{n}{2}} \sum_{\mathcal{R} \in \mathcal{P}} \Pr(\mathcal{R} \mid \mathcal{P}) \sum_{i=1}^{n-1} i(n-i)\, \mathrm{E}[\hat{\xi}_i \mid r] \\ &= \frac{\theta}{\binom{n}{2}} \sum_{r \in \mathcal{R}(\mathcal{P})} \Pr(r \mid \mathcal{P}) \sum_{i=1}^{n-1} i(n-i) \sum_{\omega \in \psi(r,i)} \frac{1}{|\omega| \binom{|\mathcal{R}|}{|w|}} \end{aligned} \tag{A.6}$$

A third estimator of θ is $\hat{\xi}_1$, the number of singletons in the sample. The conditional expectation of $\hat{\xi}_1$ given a pedigree $\mathcal{P}$ is

$$\mathrm{E}\left[\hat{\xi}_1 \mid \mathcal{P}\right] = \theta \sum_{r \in \mathcal{R}(\mathcal{P})} \frac{|\psi(r,1)|}{|\mathcal{R}(\mathcal{P})|} \Pr(r \mid \mathcal{P}), \tag{A.7}$$

since only those mutations that occur on lineages that have not coalesced with any other lineages in the early pedigree can produce singletons.

To validate these calculations, we performed simulations of 200 loci sampled from individuals whose pedigree includes some degree of IBD or inbreeding. The simulations confirm the calculated biases for the different estimators of θ (Fig. S5).

B

Pedigrees and biased estimators of M

In a constant-sized structured population with two demes and a constant rate of migration $M = 4Nm$ between demes, the expected within-deme and between-deme pairwise coalescence times are

$$\mathrm{E}[T_w] = 2 \tag{B.1}$$

$$\mathrm{E}[T_b] = 2 + 1/M, \tag{B.2}$$

Let π_w and π_b be the within-deme and between-deme mean pairwise diversity, respectively. Since $\mathrm{E}[\pi_w] = \theta\,\mathrm{E}[T_w]$ and $\mathrm{E}[\pi_b] = \theta\,\mathrm{E}[T_b]$, one estimator of M is

$$\hat{M} = \frac{\hat{\pi}_w}{2(\hat{\pi}_b - \hat{\pi}_w)}. \tag{B.3}$$

If some individuals in the sample are recently admixed, $\hat{M}$ will be biased. In general it is not possible to calculate $\mathrm{E}[\hat{M}]$, but it can be approximated by

$$\mathrm{E}[\hat{M}] \approx \frac{\mathrm{E}[\hat{\pi}_w]}{2(\mathrm{E}[\hat{\pi}_b] - \mathrm{E}[\hat{\pi}_w])}. \tag{B.4}$$

We sample two sequences from each of n_1 individuals from deme 1 and n_2 individuals from deme 2, defining $n = n_1 + n_2$ as the total (diploid) sample size. The sample is again indexed by $\mathcal{I}_n = \{1^{\mathrm{m}}, 1^{\mathrm{p}}, \ldots, n^{\mathrm{m}}, n^{\mathrm{p}}\}$, and we assume that the first $2n_1$ of these indices correspond to sequences sampled from deme 1 and the last $2n_2$ from deme 2.

In the context of a two-deme population, each group in the partitioned sample $r \in \mathcal{R}(\mathcal{P})$ is labeled 1 or 2 to indicate which deme the lineage is found in after segregation back in time through the recent sample pedigree. For two-deme reconfiguration r, let $d(r, i, j)$, $i, j \in \{1, 2\}$, be a function that gives the number of lineages originally sampled from deme i that are found in deme j after segregation through the recent sample pedigree.

Assume that the recent sample pedigree contains admixture but no IBD. In this case, we can write the expectations of $\hat{\pi}_w$ and $\hat{\pi}_b$ conditional on reconfiguration r as

$$
\begin{aligned}
\mathrm{E}\left[\hat{\pi}_w \mid r\right] = \frac{1}{\binom{2n_1}{2} + \binom{2n_2}{2}} \times \\
&\Bigg\{ \theta\, \mathrm{E}[T_w] \sum_{i=1}^{2} \sum_{j=1}^{2} \binom{d(r,i,j)}{2} + \\
&\theta\, \mathrm{E}[T_b] \left[d(r,1,1) d(r,1,2) + d(r,2,2) d(r,2,1) \right] \Bigg\}
\end{aligned}
\tag{B.5}
$$

and

$$
\begin{aligned}
\mathrm{E}\left[\hat{\pi}_b \mid r\right] = \frac{1}{4 n_1 n_2} \times \\
&\Bigg\{ \theta\, \mathrm{E}[T_b] \big(d(r,1,1) d(r,2,2) + d(r,1,2) d(r,2,1) \big) + \\
&\theta\, \mathrm{E}[T_w] \left[d(r,1,2) d(r,2,2) + d(r,2,1) d(r,1,1) \right] \Bigg\}.
\end{aligned}
\tag{B.6}
$$

The approximate expectation of $\hat{M}$ conditional on a pedigree $\mathcal{P}$ can be calculated using (B.5) and (B.6) together with

$$
\mathrm{E}[\hat{\pi}_w \mid \mathcal{P}] = \sum_{r \in \mathcal{R}(\mathcal{P})} \mathrm{E}[\hat{\pi}_w \mid r] \Pr(r \mid \mathcal{P})
$$

and

$$
\mathrm{E}[\hat{\pi}_b \mid \mathcal{P}] = \sum_{r \in \mathcal{R}(\mathcal{P})} \mathrm{E}[\hat{\pi}_b \mid r] \Pr(r \mid \mathcal{P}).
$$

Simulations of infinite-sites loci taken from samples with a single admixed ancestor confirm these calculations (Fig. S6). This method of approximating $\mathrm{E}[\hat{M} \mid \mathcal{P}]$ could be extended to accommodate recent sample pedigrees that contain both IBD and admixture, but we do not pursue this here.

C

Pairwise ARG is ergodic

Here we show that the pairwise ARG is sequentially ergodic. Let $\{t(x)\}_{x\geq 0}$ represent the random pairwise coalescence time at point x along two aligned, continuous, infinitely-long chromosomes modeled by the ARG. Let time be scaled such that the marginal distribution of $t(x)$ is exponential with rate 1 for all $x \geq 0$, and thus $\mathrm{E}[t(x)] = 1$. Let the distance across the chromosome be measured such that a segment of length l recombines apart back in time at rate $l/2$. (Equivalently, a recombination event happens in the chromosome interval $(x, x+dx)$ in the time interval $(t, t+dt)$ with infinitesimal probability $dx\,dt$.)

One useful property of $t(x)$ is that it is strongly stationary. That is, the joint distribution of $\{t(x)\}_{a\leq x\leq b}$ is the same as the joint distribution of $\{t(x)\}_{a+h\leq x\leq b+h}$ for all

$0 \leq a < b$ and $h > 0$. To see that this is the case, consider the Wiuf and Hein (1999) algorithm for constructing an ARG sequentially across the chromosome: at a given point, a genealogy is drawn from the marginal distribution of genealogies, and then the algorithm proceeds across the chromosome generating recombination events and genealogies, where at each point along the chromosome, such events are drawn from the conditional distribution given all previous coalescence and recombination events. The initial point from which the marginal genealogy is drawn has no effect on the resulting joint distribution of genealogies.

The pairwise ARG is defined to be ergodic if

$$\lim_{L\to\infty} \mathrm{E}\left[\left(\frac{1}{L}\int_0^L t(x)dx - \mathrm{E}[t(0)]\right)^2\right] = 0. \tag{C.1}$$

For a stationary process with covariance function $r(x)$, to demonstrate ergodicity it is sufficient to show that

$$\lim_{L\to\infty} \frac{1}{L}\int_0^L r(x)dx = 0. \tag{C.2}$$

(See Itō [94].) Condition (C.2) is met if $\lim_{x\to\infty} r(x) = 0$. This follows from the fact that for $K < L$,

$$\frac{1}{L}\int_0^L r(x)dx = \frac{1}{L}\int_0^K r(x)dx + \frac{1}{L}\int_K^L r(x)dx. \tag{C.3}$$

The first term of the right-hand side disappears in the $L \to \infty$ limit if $r(x)$ is bounded, and the second term can be made smaller than any arbitrary $\epsilon > 0$ by choosing a sufficiently large K [cf. 95, p. 478].

Under the ARG, the covariance between $t(0)$ and $t(x)$, $x > 0$, is

$$r(x) = \frac{x + 18}{x^2 + 13x + 18}. \tag{C.4}$$

Since $\lim_{x\to\infty} r(x) = 0$, condition (C.2) is met and the pairwise ARG is thus sequentially ergodic: the mean coalescence time across a long chromosome converges to the mean coalescence time at a single point. A similar proof could be given for the discrete-locus ARG with evenly spaced loci, which has a covariance function of the same form as the continuous-chromosome ARG. In this case, the integrals would be replaced by the corresponding sums.

D

Distribution of IBD segment lengths under the SMC' model of coalescence and recombination

Let G be the Meijer G-function, $\psi^a(b) = \frac{d^{a+1}}{db^{a+1}} \log(\Gamma(b))$ be the polygamma function, and $\Gamma(a, b) = \int_b^\infty t^{a-1} e^{-t} dt$ be the incomplete gamma function. Note that $\log(-x) = i\pi + \log(x)$ for $x > 0$. The function $f_L(l)$ gives the density of lengths of IBD segments and returns real values. Then

$$
\begin{aligned}
f_L(l) = {}& 2^{\frac{l}{2}-4} e^{-l/4} (-l)^{-\frac{l}{4}-\frac{5}{2}} l \Bigg\{ 8l\, G^{4,0}_{3,4}\left(-\frac{l}{4}\middle| \begin{matrix} 2,2,2 \\ 1,1,1,\frac{l+6}{4} \end{matrix}\right) - 8l\, G^{3,0}_{2,3}\left(-\frac{l}{4}\middle| \begin{matrix} 2,2 \\ 1,1,\frac{l+6}{4} \end{matrix}\right) \\
& - 32\, G^{3,0}_{2,3}\left(-\frac{l}{4}\middle| \begin{matrix} 2,2 \\ 1,1,\frac{l+10}{4} \end{matrix}\right) - l^2 \Gamma\left(\frac{l+2}{4}, -\frac{l}{4}\right) \\
& + l^2 \Gamma\left(\frac{l+2}{4}\right) \Bigg[4\left(l^2+l+3\right) + l\log(-l/4)\left(4l + l\log(-l/4) + 4\right) + \\
& l\Bigg[l\psi^{(0)}\left(\frac{l+2}{4}\right)^2 + l\psi^{(1)}\left(\frac{l+2}{4}\right) - 2\left(2l + l\log(-l/4) + 2\right)\psi^{(0)}\left(\frac{l+2}{4}\right)\Bigg]\Bigg] \\
& - 8l\Gamma\left(\frac{l+6}{4}, -\frac{l}{4}\right) - 16\Gamma\left(\frac{l+10}{4}, -\frac{l}{4}\right)\Bigg\}.
\end{aligned}
$$

E

Integral of cumulative coalescent rate under piecewise constant population size

There are several integrals of the form $\int_x^y e^{-k\Omega(u,y)}du$ in Equation (4.3). Assuming a piecewise population size history, we present a derivation of the solution to this integral in several steps.

$$\int_x^y e^{-k\Omega(u,y)}du = \int_x^{T_{\alpha(x)+1}} e^{-k\Omega(u,y)}du + \sum_{i=\alpha(x)+1}^{\alpha(y)-1} \int_{T_i}^{T_i+1} e^{-k\Omega(u,y)}du + \int_{T_{\alpha(y)}}^{y} e^{-k\Omega(u,y)}du$$

$$= \int_x^{T_{\alpha(x)+1}} \exp\left(-k\left[\left(T_{\alpha(u)+1}-u\right)\frac{1}{\lambda_{\alpha(u)}} + \sum_{j=\alpha(u)+1}^{\alpha(y)-1}\frac{\Delta_j}{\lambda_j} + (y-T_{\alpha(y)})\frac{1}{\lambda_{\alpha(y)}}\right]\right)du+$$

$$\sum_{i=\alpha(x)+1}^{\alpha(y)-1}\int_{T_i}^{T_{i+1}} \exp\left(-k\left[\left(T_{\alpha(u)+1}-u\right)\frac{1}{\lambda_{\alpha(u)}} + \sum_{j=\alpha(u)+1}^{\alpha(y)-1}\frac{\Delta_j}{\lambda_j} + (y-T_{\alpha(y)})\frac{1}{\lambda_{\alpha(y)}}\right]\right)du+$$

$$\int_{T_{\alpha(y)}}^{y} \exp\left(-k(y-u)\frac{1}{\lambda_{\alpha(u)}}\right)du$$

$$= \int_x^{T_{\alpha(x)+1}} \exp\left(-k\left[\left(T_{\alpha(x)+1}-u\right)\frac{1}{\lambda_{\alpha(x)}} + \sum_{j=\alpha(x)+1}^{\alpha(y)-1}\frac{\Delta_j}{\lambda_j} + (y-T_{\alpha(y)})\frac{1}{\lambda_{\alpha(y)}}\right]\right)du+$$

$$\sum_{i=\alpha(x)+1}^{\alpha(y)-1}\int_{T_i}^{T_{i+1}} \exp\left(-k\left[(T_{i+1}-u)\frac{1}{\lambda_i} + \sum_{j=i+1}^{\alpha(y)-1}\frac{\Delta_j}{\lambda_j} + (y-T_{\alpha(y)})\frac{1}{\lambda_{\alpha(y)}}\right]\right)du+$$

$$\int_{T_{\alpha(y)}}^{y} \exp\left(-k(y-u)\frac{1}{\lambda_{\alpha(y)}}\right)du$$

$$= \exp\left(-k\left[\sum_{j=\alpha(x)+1}^{\alpha(y)-1}\frac{\Delta_j}{\lambda_j} + (y-T_{\alpha(y)})\frac{1}{\lambda_{\alpha(y)}}\right]\right)\int_x^{T_{\alpha(x)+1}} \exp\left(-k\left(T_{\alpha(x)+1}-u\right)\frac{1}{\lambda_{\alpha(x)}}\right)du+$$

$$\sum_{i=\alpha(x)+1}^{\alpha(y)-1} \exp\left(-k\left[\sum_{j=i+1}^{\alpha(y)-1}\frac{\Delta_j}{\lambda_j} + (y-T_{\alpha(y)})\frac{1}{\lambda_{\alpha(y)}}\right]\right)\int_{T_i}^{T_{i+1}} \exp\left(-k\left(T_{i+1}-u\right)\frac{1}{\lambda_i}\right)du+$$

$$\int_{T_{\alpha(y)}}^{y} \exp\left(-k(y-u)\frac{1}{\lambda_{\alpha(y)}}\right)du$$

$$= \exp\left(-k\left[\sum_{j=\alpha(x)+1}^{\alpha(y)-1}\frac{\Delta_j}{\lambda_j} + (y-T_{\alpha(y)})\frac{1}{\lambda_{\alpha(y)}}\right]\right)\left[1-\exp\left(-\frac{k\left(T_{\alpha(x)+1}-x\right)}{\lambda_{\alpha(x)}}\right)\right]\frac{\lambda_{\alpha(x)}}{k}+$$

$$\sum_{i=\alpha(x)+1}^{\alpha(y)-1} \exp\left(-k\left[\sum_{j=i+1}^{\alpha(y)-1}\frac{\Delta_j}{\lambda_j} + (y-T_{\alpha(y)})\frac{1}{\lambda_{\alpha(y)}}\right]\right)\left[1-\exp\left(-\frac{k\Delta_i}{\lambda_i}\right)\right]\frac{\lambda_i}{k}+$$

$$\left[1-\exp\left(-\frac{k\left(y-T_{\alpha(y)}\right)}{\lambda_{\alpha(y)}}\right)\right]\frac{\lambda_{\alpha(y)}}{k}$$

$$= e^{-k\Omega(T_{\alpha(x)+1},y)} \left[1 - e^{-\frac{k\left(T_{\alpha(x)+1}-x\right)}{\lambda_{\alpha(x)}}}\right] \frac{\lambda_{\alpha(x)}}{k} + \sum_{i=\alpha(x)+1}^{\alpha(y)-1} e^{-k\Omega(T_{i+1},y)} \left[1 - e^{-\frac{k\Delta_i}{\lambda_i}}\right] \frac{\lambda_i}{k} + \left[1 - e^{-\frac{k\left(y-T_{\alpha(y)}\right)}{\lambda_{\alpha(y)}}}\right] \frac{\lambda_{\alpha(y)}}{k} \tag{E.1}$$

F

TSMC transitions under piecewise constant population size model

The following gives the different parts of the transition kernel $q(t_3, t_2|s_3, s_2)$ under the triploid SMC' with a piecewise constant population history as described in the text.

For $t_3 = s_3; t_2 > s_2$:

$$\frac{1}{2s_2+s_3}e^{-2\Omega(s_3,s_2)}\frac{1}{\lambda_{\alpha(t_2)}}e^{-\Omega(s_2,t_2)}\left\{\sum_{i=0}^{\alpha(s_3)-1} e^{-3\Omega(T_{i+1},s_3)}\left[1-e^{-\frac{3\Delta_i}{\lambda_i}}\right]\frac{\lambda_i}{3}+\right.$$

$$\left.\left[1-e^{-\frac{3\left(s_3-T_{\alpha(s_3)}\right)}{\lambda_{\alpha(s_3)}}}\right]\frac{\lambda_{\alpha(s_3)}}{3}\right\}+\frac{2}{2s_2+s_3}\frac{1}{\lambda_{\alpha(t_2)}}e^{-\Omega(s_2,t_2)}\times$$

$$\begin{aligned}&\left\{e^{-2\Omega(T_{\alpha(s_3)+1},s_2)}\left[1-e^{-\frac{2\left(T_{\alpha(s_3)+1}-s_3\right)}{\lambda_{\alpha(s_3)}}}\right]\frac{\lambda_{\alpha(s_3)}}{2}+\sum_{i=\alpha(s_3)+1}^{\alpha(s_2)-1}e^{-2\Omega(T_{i+1},s_2)}\left[1-e^{-\frac{2\Delta_i}{\lambda_i}}\right]\frac{\lambda_i}{2}\right.\\&\qquad\left.+\left[1-e^{-\frac{2\left(s_2-T_{\alpha(s_2)}\right)}{\lambda_{\alpha(s_2)}}}\right]\frac{\lambda_{\alpha(s_2)}}{2}\right\}\end{aligned}$$

For $t_3 < s_3; t_2 = s_3$:

$$\frac{1}{2s_2+s_3}\frac{2}{\lambda_{\alpha(t_3)}}\left\{\sum_{i=0}^{\alpha(t_3)-1}e^{-3\Omega(T_{i+1},t_3)}\left[1-e^{-\frac{3\Delta_i}{\lambda_i}}\right]\frac{\lambda_i}{3}+\left[1-e^{-\frac{3\left(t_3-T_{\alpha(t_3)}\right)}{\lambda_{\alpha(t_3)}}}\right]\frac{\lambda_{\alpha(t_3)}}{3}\right\}$$

For $t_3 < s_3; t_2 = s_2$:

$$\frac{2}{2s_2+s_3}\frac{2}{\lambda_{\alpha(t_3)}}\left\{\sum_{i=0}^{\alpha(t_3)-1}e^{-3\Omega(T_{i+1},t_3)}\left[1-e^{-\frac{3\Delta_i}{\lambda_i}}\right]\frac{\lambda_i}{3}+\left[1-e^{-\frac{3\left(t_3-T_{\alpha(t_3)}\right)}{\lambda_{\alpha(t_3)}}}\right]\frac{\lambda_{\alpha(t_3)}}{3}\right\}$$

For $t_3 > s_3; t_2 = s_2$:

$$\frac{2}{2s_2+s_3}\frac{2}{\lambda_{\alpha(t_3)}}e^{-2\Omega(s_3,t_3)}\left\{\sum_{i=0}^{\alpha(s_3)-1}e^{-3\Omega(T_{i+1},s_3)}\left[1-e^{-\frac{3\Delta_i}{\lambda_i}}\right]\frac{\lambda_i}{3}+\left[1-e^{-\frac{3\left(s_3-T_{\alpha(s_3)}\right)}{\lambda_{\alpha(s_3)}}}\right]\frac{\lambda_{\alpha(s_3)}}{3}\right\}$$

For $t_3 = s_2; t_2 > s_2$:

$$\frac{2}{2s_2+s_3}e^{-2\Omega(s_3,s_2)}\frac{1}{\lambda_{\alpha(t_2)}}e^{-\Omega(s_2,t_2)}\left\{\sum_{i=0}^{\alpha(s_3)-1}e^{-3\Omega(T_{i+1},s_3)}\left[1-e^{-\frac{3\Delta_i}{\lambda_i}}\right]\frac{\lambda_i}{3}+\right.$$

$$\left.\left[1-e^{-\frac{3\left(s_3-T_{\alpha(s_3)}\right)}{\lambda_{\alpha(s_3)}}}\right]\frac{\lambda_{\alpha(s_3)}}{3}\right\}$$

For $t_3 = s_3; t_2 = s_2$:

$$\frac{1}{2s_2+s_3}\left\{s_3-\sum_{i=0}^{\alpha(s_3)-1}e^{-3\Omega(T_{i+1},s_3)}\left[1-e^{-\frac{3\Delta_i}{\lambda_i}}\right]\frac{\lambda_i}{3}-\left[1-e^{-\frac{3\left(s_3-T_{\alpha(s_3)}\right)}{\lambda_{\alpha(s_3)}}}\right]\frac{\lambda_{\alpha(s_3)}}{3}\right\}+$$

$$\frac{1}{2s_2+s_3}\frac{1}{2}\left[1-e^{-2\Omega(s_3,s_2)}\right]\left\{\sum_{i=0}^{\alpha(s_3)-1}e^{-3\Omega(T_{i+1},s_3)}\left[1-e^{-\frac{3\Delta_i}{\lambda_i}}\right]\frac{\lambda_i}{3}+\right.$$

$$\left.\left[1-e^{-\frac{3\left(s_3-T_{\alpha(s_3)}\right)}{\lambda_{\alpha(s_3)}}}\right]\frac{\lambda_{\alpha(s_3)}}{3}\right\}+$$

$$\frac{1}{2s_2+s_3}\left[s_2-s_3-e^{-2\Omega(T_{\alpha(s_3)+1},t_2)}\left(1-e^{-\frac{2\left(T_{\alpha(s_3)+1}-s_3\right)}{\lambda_{\alpha(s_3)}}}\right)\frac{\lambda_{\alpha(s_3)}}{2}-\right.$$

$$\left.\sum_{i=\alpha(s_3)+1}^{\alpha(t_2)-1}e^{-2\Omega(T_{i+1},t_2)}\left(1-e^{-\frac{2\Delta_i}{\lambda_i}}\right)\frac{\lambda_i}{2}-\left(1-e^{-\frac{2\left(t_2-T_{\alpha(t_2)}\right)}{\lambda_{\alpha(t_2)}}}\right)\frac{\lambda_{\alpha(t_2)}}{2}\right]$$

G

Discrete-time approximations to TSMC transition probabilities

Here we present the transition probabilities from state (i, j) to (k, l) in our discrete-time approximation to the continuous-time TSMC model. Multiple steps are shown for the derivation of the first part, and only the final form of the expression is shown for the other parts.

For $i = k < j < l$, $(t_3 = s_3\,; t_2 > s_2)$:

$$= \frac{1}{2s_2 + s_3} e^{-2\Omega(s_3,s_2)} \frac{1}{\lambda_{\alpha(t_2)}} e^{-\Omega(s_2,t_2)} \left\{ \sum_{i=0}^{\alpha(s_3)-1} e^{-3\Omega(T_{i+1},s_3)} \left[1 - e^{-\frac{3\Delta_i}{\lambda_i}}\right] \frac{\lambda_i}{3} + \right.$$

$$\left. \left[1 - e^{-\frac{3\left(s_3 - T_{\alpha(s_3)}\right)}{\lambda_{\alpha(s_3)}}}\right] \frac{\lambda_{\alpha(s_3)}}{3} \right\} + \frac{2}{2s_2 + s_3} \frac{1}{\lambda_{\alpha(t_2)}} e^{-\Omega(s_2,t_2)} \times$$

$$\left\{ e^{-2\Omega(T_{\alpha(s_3)+1},s_2)} \left[1 - e^{-\frac{2\left(T_{\alpha(s_3)+1} - s_3\right)}{\lambda_{\alpha(s_3)}}}\right] \frac{\lambda_{\alpha(s_3)}}{2} + \sum_{i=\alpha(s_3)+1}^{\alpha(s_2)-1} e^{-2\Omega(T_{i+1},s_2)} \left[1 - e^{-\frac{2\Delta_i}{\lambda_i}}\right] \frac{\lambda_i}{2} \right.$$

$$\left. + \left[1 - e^{-\frac{2\left(s_2 - T_{\alpha(s_2)}\right)}{\lambda_{\alpha(s_2)}}}\right] \frac{\lambda_{\alpha(s_2)}}{2} \right\}$$

$$= \int_{T_k}^{T_{k+1}} \int_{T_l}^{T_{l+1}} \frac{1}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]} e^{-2\Omega(\mathrm{E}_{i,j}[s_3],\mathrm{E}_{i,j}[s_2])} \frac{1}{\lambda_l} e^{-\Omega(\mathrm{E}_{i,j}[s_2],t_2)} \times$$

$$\left\{ \sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,j}[s_3])} \left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} + \left[1 - e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3] - T_i\right)}{\lambda_i}}\right] \frac{\lambda_i}{3} \right\} \delta(t_3 - s_3) dt_2 dt_3 +$$

$$\int_{T_k}^{T_{k+1}} \int_{T_l}^{T_{l+1}} \frac{2}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]} \frac{1}{\lambda_l} e^{-\Omega(\mathrm{E}_{i,j}[s_2],t_2)} \times$$

$$\left\{ e^{-2\Omega(T_{i+1},\mathrm{E}_{i,j}[s_2])} \left[1 - e^{-\frac{2\left(T_{i+1} - \mathrm{E}_{i,j}[s_3]\right)}{\lambda_i}}\right] \frac{\lambda_i}{2} + \sum_{a=i+1}^{j-1} e^{-2\Omega(T_{a+1},\mathrm{E}_{i,j}[s_2])} \left[1 - e^{-\frac{2\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{2} + \right.$$

$$\left. \left[1 - e^{-\frac{2\left(\mathrm{E}_{i,j}[s_2] - T_j\right)}{\lambda_j}}\right] \frac{\lambda_j}{2} \right\} \delta(t_3 - s_3) dt_2 dt_3$$

$$
\begin{aligned}
&= \int_{T_l}^{T_{l+1}} \frac{1}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]} e^{-2\Omega(\mathrm{E}_{i,j}[s_3],\mathrm{E}_{i,j}[s_2])} \frac{1}{\lambda_l} e^{-\Omega(\mathrm{E}_{i,j}[s_2],t_2)} \times \\
&\quad \left\{ \sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,j}[s_3])} \left[1-e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} + \left[1-e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3]-T_i\right)}{\lambda_i}}\right] \frac{\lambda_i}{3} \right\} dt_2 + \\
&\quad \int_{T_l}^{T_{l+1}} \frac{2}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]} \frac{1}{\lambda_l} e^{-\Omega(\mathrm{E}_{i,j}[s_2],t_2)} \times \\
&\quad \left\{ e^{-2\Omega(T_{i+1},\mathrm{E}_{i,j}[s_2])} \left[1-e^{-\frac{2\left(T_{i+1}-\mathrm{E}_{i,j}[s_3]\right)}{\lambda_i}}\right] \frac{\lambda_i}{2} + \sum_{a=i+1}^{j-1} e^{-2\Omega(T_{a+1},\mathrm{E}_{i,j}[s_2])} \left[1-e^{-\frac{2\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{2} + \right. \\
&\qquad\qquad \left. \left[1-e^{-\frac{2\left(\mathrm{E}_{i,j}[s_2]-T_j\right)}{\lambda_j}}\right] \frac{\lambda_j}{2} \right\} dt_2 \\
&= \frac{1}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]} e^{-2\Omega(\mathrm{E}_{i,j}[s_3],\mathrm{E}_{i,j}[s_2])} \frac{1}{\lambda_l} \times \\
&\quad \left\{ \sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,j}[s_3])} \left[1-e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} + \left[1-e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3]-T_i\right)}{\lambda_i}}\right] \frac{\lambda_i}{3} \right\} \times \\
&\qquad\qquad \int_{T_l}^{T_{l+1}} e^{-\Omega(\mathrm{E}_{i,j}[s_2],t_2)} dt_2 + \\
&\quad \frac{2}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]} \frac{1}{\lambda_l} \left\{ e^{-2\Omega(T_{i+1},\mathrm{E}_{i,j}[s_2])} \left[1-e^{-\frac{2\left(T_{i+1}-\mathrm{E}_{i,j}[s_3]\right)}{\lambda_i}}\right] \frac{\lambda_i}{2} + \right. \\
&\quad \left. \sum_{a=i+1}^{j-1} e^{-2\Omega(T_{a+1},\mathrm{E}_{i,j}[s_2])} \left[1-e^{-\frac{2\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{2} + \left[1-e^{-\frac{2\left(\mathrm{E}_{i,j}[s_2]-T_j\right)}{\lambda_j}}\right] \frac{\lambda_j}{2} \right\} \times \\
&\qquad\qquad \int_{T_l}^{T_{l+1}} e^{-\Omega(\mathrm{E}_{i,j}[s_2],t_2)} dt_2 \\
&= \frac{1}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]} e^{-2\Omega(\mathrm{E}_{i,j}[s_3],\mathrm{E}_{i,j}[s_2])} \frac{1}{\lambda_l} \times \\
&\quad \left\{ \sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,j}[s_3])} \left[1-e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} + \left[1-e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3]-T_i\right)}{\lambda_i}}\right] \frac{\lambda_i}{3} \right\} e^{-\Omega(\mathrm{E}_{i,j}[s_2],T_l)} \times \\
&\quad \int_{T_l}^{T_{l+1}} e^{-\Omega(T_l,t_2)} dt_2 + \frac{2}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]} \frac{1}{\lambda_l} \left\{ e^{-2\Omega(T_{i+1},\mathrm{E}_{i,j}[s_2])} \left[1-e^{-\frac{2\left(T_{i+1}-\mathrm{E}_{i,j}[s_3]\right)}{\lambda_i}}\right] \frac{\lambda_i}{2} + \right. \\
&\quad \left. \sum_{a=i+1}^{j-1} e^{-2\Omega(T_{a+1},\mathrm{E}_{i,j}[s_2])} \left[1-e^{-\frac{2\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{2} + \left[1-e^{-\frac{2\left(\mathrm{E}_{i,j}[s_2]-T_j\right)}{\lambda_j}}\right] \frac{\lambda_j}{2} \right\} e^{-\Omega(\mathrm{E}_{i,j}[s_2],T_l)} \times \\
&\quad \int_{T_l}^{T_{l+1}} e^{-\Omega(T_l,t_2)} dt_2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]} e^{-2\Omega(\mathrm{E}_{i,j}[s_3],\mathrm{E}_{i,j}[s_2])}\frac{1}{\lambda_l}\times\\
&\quad \left\{\sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,j}[s_3])}\left[1-e^{-\frac{3\Delta_a}{\lambda_a}}\right]\frac{\lambda_a}{3}+\left[1-e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3]-T_i\right)}{\lambda_i}}\right]\frac{\lambda_i}{3}\right\} e^{-\Omega(\mathrm{E}_{i,j}[s_2],T_l)}\times\\
&\quad \int_{T_l}^{T_{l+1}} e^{-\frac{t_2-T_l}{\lambda_l}}dt_2 + \frac{2}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]}\frac{1}{\lambda_l}\left\{e^{-2\Omega(T_{i+1},\mathrm{E}_{i,j}[s_2])}\left[1-e^{-\frac{2\left(T_{i+1}-\mathrm{E}_{i,j}[s_3]\right)}{\lambda_i}}\right]\frac{\lambda_i}{2}+\right.\\
&\quad \left.\sum_{a=i+1}^{j-1} e^{-2\Omega(T_{a+1},\mathrm{E}_{i,j}[s_2])}\left[1-e^{-\frac{2\Delta_a}{\lambda_a}}\right]\frac{\lambda_a}{2}+\left[1-e^{-\frac{2\left(\mathrm{E}_{i,j}[s_2]-T_j\right)}{\lambda_j}}\right]\frac{\lambda_j}{2}\right\}\times\\
&\quad e^{-\Omega(\mathrm{E}_{i,j}[s_2],T_l)}\int_{T_l}^{T_{l+1}} e^{-\frac{t_2-T_l}{\lambda_l}}dt_2\\
&= \frac{1}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]} e^{-2\Omega(\mathrm{E}_{i,j}[s_3],\mathrm{E}_{i,j}[s_2])}\frac{1}{\lambda_l}\times\\
&\quad \left\{\sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,j}[s_3])}\left[1-e^{-\frac{3\Delta_a}{\lambda_a}}\right]\frac{\lambda_a}{3}+\left[1-e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3]-T_i\right)}{\lambda_i}}\right]\frac{\lambda_i}{3}\right\} e^{-\Omega(\mathrm{E}_{i,j}[s_2],T_l)}\times\\
&\quad \lambda_l\left(1-e^{-\frac{\Delta_l}{\lambda_l}}\right)+\\
&\quad \frac{2}{2\,\mathrm{E}_{i,j}[s_2]+\mathrm{E}_{i,j}[s_3]}\frac{1}{\lambda_l}\left\{e^{-2\Omega(T_{i+1},\mathrm{E}_{i,j}[s_2])}\left[1-e^{-\frac{2\left(T_{i+1}-\mathrm{E}_{i,j}[s_3]\right)}{\lambda_i}}\right]\frac{\lambda_i}{2}+\right.\\
&\quad \left.\sum_{a=i+1}^{j-1} e^{-2\Omega(T_{a+1},\mathrm{E}_{i,j}[s_2])}\left[1-e^{-\frac{2\Delta_a}{\lambda_a}}\right]\frac{\lambda_a}{2}+\left[1-e^{-\frac{2\left(\mathrm{E}_{i,j}[s_2]-T_j\right)}{\lambda_j}}\right]\frac{\lambda_j}{2}\right\}\times\\
&\quad e^{-\Omega(\mathrm{E}_{i,j}[s_2],T_l)}\lambda_l\left(1-e^{-\frac{\Delta_l}{\lambda_l}}\right)
\end{aligned}
\tag{G.1}
$$

Here we assume that $\Delta_n = \infty$ and thus $e^{-\Delta_n/\lambda_n} = 0$.

For $i = k < l < j$ $(t_3 = s_3; t_2 < s_2)$:

$$\begin{aligned}
= &\frac{1}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]}\frac{1}{\lambda_l} \\
&\left\{\sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,j}[s_3])}\left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right]\frac{\lambda_a}{3} + \left[1 - e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3] - T_i\right)}{\lambda_i}}\right]\frac{\lambda_i}{3}\right\}\times \\
&e^{-2\Omega(\mathrm{E}_{i,j}[s_3],T_l)}\frac{\lambda_l}{2}\left(1 - e^{-\frac{2\Delta_l}{\lambda_l}}\right) + \\
&\frac{2}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]}\frac{1}{\lambda_l}\left(\left[1 - e^{-\frac{2\left(T_{i+1} - \mathrm{E}_{i,j}[s_3]\right)}{\lambda_i}}\right]\frac{\lambda_i}{2}\right)e^{-2\Omega(T_{i+1},T_l)}\frac{\lambda_l}{2}\left(1 - e^{-\frac{2\Delta_l}{\lambda_l}}\right) + \\
&\frac{2}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]}\frac{1}{\lambda_l}\left[\sum_{a=i+1}^{l-1}\left[1 - e^{-\frac{2\Delta_a}{\lambda_a}}\right]\frac{\lambda_a}{2}e^{-2\Omega(T_{a+1},T_l)}\frac{\lambda_l}{2}\left(1 - e^{-\frac{2\Delta_l}{\lambda_l}}\right)\right] + \\
&\frac{2}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]}\frac{1}{\lambda_l}\frac{\lambda_l}{2}\left[\Delta_l - \frac{\lambda_l}{2}\left(1 - e^{-\frac{2\Delta_l}{\lambda_l}}\right)\right]
\end{aligned} \tag{G.2}$$

For $k < i = l < j$ $(t_3 < s_3; t_2 = s_3)$:

$$\begin{aligned}
= &\frac{1}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]}\frac{2}{\lambda_k}\left\{\frac{\lambda_k}{3}\left[1 - e^{-\frac{3\Delta_k}{\lambda_k}}\right]\sum_{a=0}^{k-1}\left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right]\frac{\lambda_a}{3}e^{-3\Omega(T_{a+1},T_k)} + \right. \\
&\left.\frac{\lambda_k}{3}\left[\Delta_k - \frac{\lambda_k}{3}\left(1 - e^{-\frac{3\Delta_k}{\lambda_k}}\right)\right]\right\}
\end{aligned} \tag{G.3}$$

For $k < i < j = l$ $(t_3 < s_3; t_2 = s_2)$:

$$\begin{aligned}
= &\frac{2}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]}\frac{2}{\lambda_k}\left\{\frac{\lambda_k}{3}\left[1 - e^{-\frac{3\Delta_k}{\lambda_k}}\right]\sum_{a=0}^{k-1}\left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right]\frac{\lambda_a}{3}e^{-3\Omega(T_{a+1},T_k)} + \right. \\
&\left.\frac{\lambda_k}{3}\left[\Delta_k - \frac{\lambda_k}{3}\left(1 - e^{-\frac{3\Delta_k}{\lambda_k}}\right)\right]\right\}
\end{aligned} \tag{G.4}$$

For $i < k < j = l$ $(t_3 > s_3; t_2 = s_2)$:

$$\begin{aligned}
&= \frac{2}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]} \frac{2}{\lambda_k} \times \\
&\quad \left\{ \sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1}, \mathrm{E}_{i,j}[s_3])} \left[1 - e^{-\frac{3\Delta_a}{\lambda_a}} \right] \frac{\lambda_a}{3} + \left[1 - e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3] - T_i\right)}{\lambda_i}} \right] \frac{\lambda_i}{3} \right\} \times \\
&\quad e^{-2\Omega(\mathrm{E}_{i,j}[s_3], T_k)} \frac{\lambda_k}{2} \left[1 - e^{-\frac{2\Delta_k}{\lambda_k}} \right]
\end{aligned} \tag{G.5}$$

For $i < j = k < l$ $(t_3 = s_2; t_2 > s_2)$:

$$\begin{aligned}
&= \frac{2}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]} e^{-2\Omega(\mathrm{E}_{i,j}[s_3], \mathrm{E}_{i,j}[s_2])} \frac{1}{\lambda_l} \times \\
&\quad \left\{ \sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1}, \mathrm{E}_{i,j}[s_3])} \left[1 - e^{-\frac{3\Delta_a}{\lambda_a}} \right] \frac{\lambda_a}{3} + \left[1 - e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3] - T_i\right)}{\lambda_i}} \right] \frac{\lambda_i}{3} \right\} \times \\
&\quad e^{-\Omega(\mathrm{E}_{i,j}[s_2], T_l)} \lambda_l \left[1 - e^{-\frac{\Delta_l}{\lambda_l}} \right]
\end{aligned} \tag{G.6}$$

Here again we assume $\Delta_n = \infty$ and thus $e^{-\Delta_n/\lambda_n} = 0$.

The following cases require special consideration. For $i = k = l < j$ and $t_3 = s_3; t_2 < s_2$. (We also need to consider $t_3 < s_3; t_2 = s_3$, and the total probability for this discrete transition will be the sum of these.)

$$\begin{aligned}
&= \frac{1}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]} \Bigg\{ \frac{1}{\lambda_k} \times \\
&\quad \left[\sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1}, \mathrm{E}_{i,j}[s_3])} \left(1 - e^{-\frac{3\Delta_a}{\lambda_a}} \right) \frac{\lambda_a}{3} + \left(1 - e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3] - T_i\right)}{\lambda_i}} \right) \frac{\lambda_i}{3} \right] \times \\
&\quad \frac{\lambda_k}{2} \left[1 - e^{-\frac{2(T_{k+1} - \mathrm{E}_{i,j}[s_3])}{\lambda_k}} \right] + T_{k+1} - \mathrm{E}_{i,j}[s_3] - \frac{\lambda_k}{2} \left[1 - e^{-\frac{2(T_{k+1} - \mathrm{E}_{i,j}[s_3])}{\lambda_k}} \right] \Bigg\}
\end{aligned} \tag{G.7}$$

For $i = k = l < j$ and $t_3 < s_3; t_2 = s_3$:

$$\begin{aligned} = \frac{1}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]} \frac{2}{\lambda_k} \Bigg\{ &\sum_{a=0}^{k-1} \left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} e^{-3\Omega(T_{a+1},T_k)} \frac{\lambda_k}{3} \left[1 - e^{-\frac{3(\mathrm{E}_{i,j}[s_3]-T_k)}{\lambda_k}}\right] + \\ &\frac{\lambda_k}{3}\left(\mathrm{E}_{i,j}[s_3] - T_k - \frac{\lambda_k}{3}\left[1 - e^{-\frac{3(\mathrm{E}_{i,j}[s_3]-T_k)}{\lambda_k}}\right]\right) \Bigg\} \end{aligned} \tag{G.8}$$

Another case that requires special consideration is $i < k = l = j$. For $i < k = l = j$ and $t_3 > s_3\,; t_2 = s_2$:

$$\begin{aligned} = &\frac{2}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]} \frac{1}{\lambda_k} e^{-2\Omega(\mathrm{E}_{i,j}[s_3],\mathrm{E}_{i,j}[s_2])} \times \\ &\left(\sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,j}[s_3])} \left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} + \left[1 - e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3]-T_i\right)}{\lambda_i}}\right] \frac{\lambda_i}{3} \right) \times \\ &\lambda_k \left[1 - \delta(k-n) e^{-\frac{T_{k+1}-\mathrm{E}_{i,j}[s_2]}{\lambda_k}}\right] \end{aligned} \tag{G.9}$$

The delta function is a way to ensure correctness when $k = l = j = n$.

When $i < k = l = j$ and $t_3 = s_2\,; t_2 > s_2$:

$$\begin{aligned} = &\frac{2}{2\,\mathrm{E}_{i,j}[s_2] + \mathrm{E}_{i,j}[s_3]} \frac{1}{\lambda_k} e^{-2\Omega(\mathrm{E}_{i,j}[s_3],\mathrm{E}_{i,j}[s_2])} \times \\ &\left(\sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,j}[s_3])} \left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} + \left[1 - e^{-\frac{3\left(\mathrm{E}_{i,j}[s_3]-T_i\right)}{\lambda_i}}\right] \frac{\lambda_i}{3} \right) \times \\ &\lambda_k \left[1 - \delta(k-n) e^{-\frac{T_{k+1}-\mathrm{E}_{i,j}[s_2]}{\lambda_k}}\right] \end{aligned} \tag{G.10}$$

For $i = j = k < l$ and $t_3 = s_3; t_2 > s_2$:

$$\begin{aligned}
&= \frac{1}{2\,\mathrm{E}_{i,i}[s_2] + \mathrm{E}_{i,i}[s_3]} \times \\
&\Bigg\{ e^{-2\Omega(\mathrm{E}_{i,i}[s_3],\mathrm{E}_{i,i}[s_2])} \frac{1}{\lambda_l} \left[\sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,i}[s_3])} \left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} + \right. \\
&\left. \left[1 - e^{-\frac{3\left(\mathrm{E}_{i,i}[s_3]-T_i\right)}{\lambda_i}}\right] \frac{\lambda_i}{3} \right] e^{-\Omega(\mathrm{E}_{i,i}[s_2],T_l)} \lambda_l \left[1 - \delta(l-n) e^{-\frac{\Delta_l}{\lambda_l}}\right] + \\
&\frac{2}{\lambda_l} \frac{\lambda_i}{2} \left[1 - e^{-\frac{2(\mathrm{E}_{i,i}[s_2]-\mathrm{E}_{i,i}[s_3])}{\lambda_i}}\right] e^{-\Omega(\mathrm{E}_{i,i}[s_2],T_l)} \lambda_l \left[1 - \delta(l-n) e^{-\frac{\Delta_l}{\lambda_l}}\right] \Bigg\}
\end{aligned} \tag{G.11}$$

For $i = j = k < l$ and $t_3 = s_2; t_2 > s_2$:

$$\begin{aligned}
&= \frac{2}{2\,\mathrm{E}_{i,i}[s_2] + \mathrm{E}_{i,i}[s_3]} e^{-2\Omega(\mathrm{E}_{i,i}[s_3],\mathrm{E}_{i,i}[s_2])} \frac{1}{\lambda_l} \left(e^{-\Omega(\mathrm{E}_{i,i}[s_2],T_l)} \lambda_l \left[1 - e^{-\delta(l-n)\frac{\Delta_l}{\lambda_l}}\right] \right) \times \\
&\left(\sum_{a=0}^{i-1} e^{-3\Omega(T_{a+1},\mathrm{E}_{i,i}[s_3])} \left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} + \left[1 - e^{-\frac{3\left(\mathrm{E}_{i,i}[s_3]-T_i\right)}{\lambda_i}}\right] \frac{\lambda_i}{3} \right)
\end{aligned} \tag{G.12}$$

Another case that requires special consideration: $k < i = j = l$. This includes either $t_3 < s_3; t_2 = s_2$ or $t_3 < s_3; t_2 = s_3$. For $k < i = j = l$ and $t_3 < s_3; t_2 = s_2$:

$$\begin{aligned}
&= \frac{2}{2\,\mathrm{E}_{i,i}[s_2] + \mathrm{E}_{i,i}[s_3]} \frac{2}{\lambda_k} \Bigg\{ \frac{\lambda_k}{3} \left(1 - e^{-\frac{3\Delta_k}{\lambda_k}}\right) \times \\
&\left(\sum_{a=0}^{k-1} \left[1 - e^{-\frac{3\Delta_a}{\lambda_a}}\right] \frac{\lambda_a}{3} e^{-3\Omega(T_{a+1},T_k)} \right) + \frac{\lambda_k}{3} \left[\Delta_k - \frac{\lambda_k}{3} \left(1 - e^{-\frac{3\Delta_k}{\lambda_k}}\right) \right] \Bigg\}
\end{aligned} \tag{G.13}$$

For $k < i = j = l$ and $t_3 < s_3; t_2 = s_3$:

$$
\begin{aligned}
&= \frac{2}{2s_2+s_3}\frac{1}{\lambda_k}\left(\sum_{a=0}^{k-1}\left[1-e^{-\frac{3\Delta_a}{\lambda_a}}\right]\frac{\lambda_a}{3}e^{-3\Omega(T_{a+1},T_k)}\frac{\lambda_k}{3}\left[1-e^{-\frac{3\Delta_k}{\lambda_k}}\right]+\right.\\
&\qquad\left.\frac{\lambda_k}{3}\left[\Delta_k-\frac{\lambda_k}{3}\left(1-e^{-\frac{3\Delta_k}{\lambda_k}}\right)\right]\right)
\end{aligned}
\tag{G.14}
$$

The diagonal case of $k = i\,; l = j$ is calculated by subtracting the sum of the off-diagonal entries from unity.

H

TSMC transition functions for a triploid asexual lineage that was previously diploid

Here we consider the possibility that the asexual triploid lineages were first diploid sexual lineages and were subsequently fertilized by a a haploid sexual sperm, making a triploid asexual lineage. The goal is to infer the time between this diploid asexual phase and the triploid asexual phase, along with everything else that was considered in inference previously. In this case, the state at each position along the genome is the vector (t_3, t_2, W), where $W \in \{0,1\}$ is an indicator for whether the branch introduced by the triploidizing sperm subtends one of the diploid sexual branches coalescing at t_3 $(W = 1)$ or t_2 $(W = 0)$. The time at transition to diploid asexuality is defined as $D_2 = 0$, and the transition to triploid asexuality is defined as $D_3 < 0$. See Fig. H.1 for an illustration of these states.

Having to keep track of W doubles the number of states, and since the complexity of the forward-backward algorithm is squared in the number of states, the runtime should increase by ~ 4.

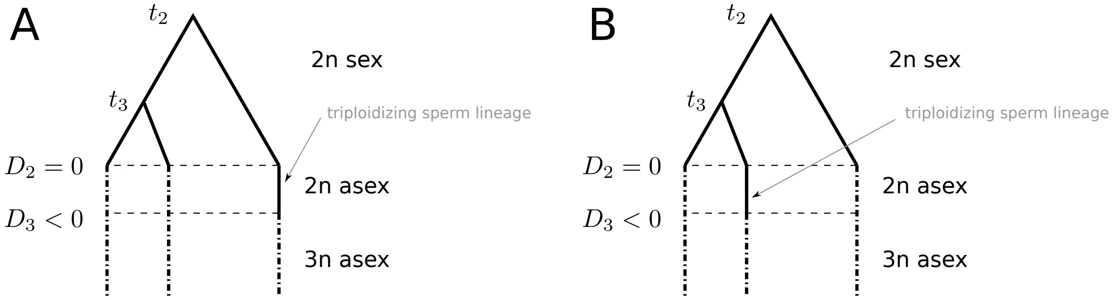

Figure H.1: Illustration of TSMC states when modeling first a transition to diploid asexual reproduction and then a transition to triploid asexual reproduction. Two states with the same coalescence times t_3 and t_2 but with the triploidizing sperm's lineage subtending a branch that coalesces with the others (in the diploid sexual phase) at t_2 (panel A, $W = 0$) and at t_3 (panel B, $W = 1$). $D_2 = 0$ is the time of transition to diploid asexuality, and $D_3 < 0$ is the time of the transition to triploidy.

The transition probabilities of the hidden model largely remain the same. A few changes occur:

1. In all previous probabilities, the factor $(2t_2 + t_3)^{-1}$ is replaced by the reciprocal of the new total sexual tree length, $(2t_2 + t_3 - D_3)^{-1}$, recalling that $D_3 < 0$. These factors are included unchanged all the way through the derivations and are still present in the final discrete transition probabilities, so they can just be replaced with the new factor.

2. Certain transitions will now change W as well. For example, if $W = 0$, (recall, the sexual lineage joins a branch that coalesces at t_2), the transition $(s_3, s_2) \to (t_3 < s_3, t_2 = s_3)$ implies that now $W = 1$.

3. Additional recombination probability must be included, arising from recombination events that occur on the triploidizing sperm's sexual lineage between $t = D_3$ and $t = 0$.

To do this properly, it will also now be necessary to model the population size changes that occur in the sexual population during the diploid asexual phase, since the triploidizing sperm's lineage will experience those demographic changes. However, the only probabilities that depend on these population sizes are the "healing" probabilities specific to the SMC', meaning that only the "effective recombination rate" for this single lineage will depend on this part of the demographic history. For this reason, we will assume that the diploid sexual population containing the triploidizing sperm is constant in size. This approximation would be unnecessary under the SMC model (vs. the SMC'), since healing is impossible under the SMC.

In general, we can expect there to be very little information about the length of this hypothesized diploid asexual interval in the triploid asexual's present-day genome. Including this additional period of diploid asexuality doesn't change the emission probabilities or the equilibrium distribution of (t_3, t_2) at all; it changes only the transition probabilities of the hidden model by affecting the probability of recombination.

The following are the "supplemental" probabilities of transition at the site of a recombination event, due to recombination events happening along the lineage of the triploidizing sperm in the time interval (D_3, D_2).

For $t_3 = s_3; t_2 > s_2$, $W = 0$, and $W' = 0$:

$$\int_{D_3}^{0} \frac{du}{2s_2 + s_3 - D_3} e^{-\Omega(u,0)} e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,s_2)} \frac{1}{\lambda(t_2)} e^{-\Omega(s_2,t_2)}$$

For $t_3 = s_3; t_2 < s_2$, $W = 0$, and $W' = 0$:

$$\int_{D_3}^{0} \frac{du}{2s_2 + s_3 - D_3} e^{-\Omega(u,0)} e^{-3\Omega(0,s_3)} \frac{1}{\lambda(t_2)} e^{-2\Omega(s_3,t_2)}$$

For $t_3 < s_3; t_2 = s_3$, $W = 0$, and $W' = 1$:

$$\int_{D_3}^{0} \frac{du}{2s_2 + s_3 - D_3} e^{-\Omega(u,0)} \frac{2}{\lambda(t_3)} e^{-3\Omega(0,t_3)}$$

For $t_3 < s_3; t_2 = s_2$, $W = 1$, and $W' = 1$:

$$\int_{D_3}^{0} \frac{du}{2s_2 + s_3 - D_3} e^{-\Omega(u,0)} e^{-3\Omega(0,t_3)} \frac{2}{\lambda(t_3)}$$

For $t_3 > s_3; t_2 = s_2$, $W = 1$, $W' = 1$:

$$\int_{D_3}^{0} \frac{du}{2s_2 + s_3 - D_3} e^{-\Omega(u,0)} e^{-3\Omega(0,s_3)} \frac{2}{\lambda(t_3)} e^{-2\Omega(s_3,t_3)}$$

For $t_3 = s_2; t_2 > s_2$, $W = 1$, and $W' = 0$:

$$\int_{D_3}^{0} \frac{du}{2s_2 + s_3 - D_3} e^{-\Omega(u,0)} e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,s_2)} \frac{1}{\lambda(t_2)} e^{-\Omega(s_2,t_2)}$$

Let the relative population size between D_3 and D_2 be λ_d. As before, replace s_3 with $\mathrm{E}_{i,j}[s_3]$ and s_2 with $\mathrm{E}_{i,j}[s_2]$. The following gives the different parts of the discrete transition probabilities with this extended state space. In each, $q(i,j,k,l)$ is the previous transition probability, calculated without the extension of the state space to both a transition to diploid asexual reproduction and a transition to triploid asexual reproduction. We provide the several steps for derivation of the final expression for the first part, and for the other parts we provide only the final expression.

For $i = k < j < l$, $(t_3 = s_3\,; t_2 > s_2)$, $W' = W$:

$$
\begin{aligned}
&= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q(i, j, k, l) + \\
&\quad \delta(W) \frac{1}{2s_2 + s_3 - D_3} \lambda_d \left(1 - e^{\frac{D_3}{\lambda_d}}\right) e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,s_2)} \frac{1}{\lambda_l} \int_{T_l}^{T_{l+1}} e^{-\Omega(s_2,t_2)} dt_2 \\
&= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q(i, j, k, l) + \\
&\quad \delta(W) \frac{1}{2s_2 + s_3 - D_3} \lambda_d \left(1 - e^{\frac{D_3}{\lambda_d}}\right) e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,s_2)} \frac{1}{\lambda_l} e^{-\Omega(s_2,T_l)} \int_{T_l}^{T_{l+1}} e^{-\Omega(T_l,t_2)} dt_2 \\
&= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q(i, j, k, l) + \\
&\quad \delta(W) \frac{1}{2s_2 + s_3 - D_3} \lambda_d \left(1 - e^{\frac{D_3}{\lambda_d}}\right) e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,s_2)} \frac{1}{\lambda_l} e^{-\Omega(s_2,T_l)} \lambda_l \left(1 - e^{-\frac{\Delta_l}{\lambda_l}}\right) \\
&= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q(i, j, k, l) + \\
&\quad \delta(W) \frac{1}{2s_2 + s_3 - D_3} \lambda_d \left(1 - e^{\frac{D_3}{\lambda_d}}\right) e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,s_2)} e^{-\Omega(s_2,T_l)} \left(1 - e^{-\frac{\Delta_l}{\lambda_l}}\right)
\end{aligned}
\tag{H.1}
$$

As above we assume that $\Delta_n = \infty$ and thus $e^{-\Delta_n/\lambda_n} = 0$.

For $i = k < l < j$ $(t_3 = s_3; t_2 < s_2)$ and $W = W'$:

$$
\begin{aligned}
&= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q(i, j, k, l) + \\
&\delta(W) \frac{1}{2s_2 + s_3 - D_3} \lambda_d \left(1 - e^{\frac{D_3}{\lambda_d}}\right) e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,T_l)} \frac{1}{2} \left(1 - e^{-\frac{2\Delta_l}{\lambda_l}}\right)
\end{aligned}
\tag{H.2}
$$

For $k < i = l < j$ $(t_3 < s_3; t_2 = s_3)$, $W = 0$, $W' = 1$:

$$\begin{aligned} &= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q(i,j,k,l) + \\ &\quad \frac{\lambda_d \left(1 - e^{\frac{D_3}{\lambda_d}}\right)}{2s_2 + s_3 - D_3} \frac{2}{3} e^{-3\Omega(0,T_k)} \left(1 - e^{-\frac{3\Delta_k}{\lambda_k}}\right) \end{aligned} \tag{H.3}$$

For $k < i = l < j$ $(t_3 < s_3; t_2 = s_3)$, $W = 1$, $W' \in \{0, 1\}$:

$$= \frac{1}{2} \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q(i,j,k,l) \tag{H.4}$$

For $k < i < j = l$ $(t_3 < s_3; t_2 = s_2)$ and $W' = W$:

$$\begin{aligned} &= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q(i,j,k,l) + \\ &\quad \delta(W-1) \frac{\lambda_d \left(1 - e^{\frac{D_3}{\lambda_d}}\right)}{2s_2 + s_3 - D_3} \frac{2}{3} e^{-3\Omega(0,T_k)} \left(1 - e^{-\frac{3\Delta_k}{\lambda_k}}\right) \end{aligned} \tag{H.5}$$

For $i < k < j = l$ $(t_3 > s_3; t_2 = s_2)$ and $W' = W$:

$$\begin{aligned} &= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q(i,j,k,l) + \\ &\quad \delta(W-1) \frac{\lambda_d \left(1 - e^{\frac{D_3}{\lambda_d}}\right)}{2s_2 + s_3 - D_3} e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,T_k)} \left(1 - e^{-\frac{2\Delta_k}{\lambda_k}}\right) \end{aligned} \tag{H.6}$$

For $i < j = k < l$ ($t_3 = s_2; t_2 > s_2$) and $W = 1$:

$$\begin{aligned}
&= \frac{2s_2+s_3}{2s_2+s_3-D_3}\frac{1}{2}\,q(i,j,k,l)+ \\
&\quad \delta(W')\frac{\lambda_d\left(1-e^{\frac{D_3}{\lambda_d}}\right)}{2s_2+s_3-D_3}e^{-3\Omega(0,s_3)}e^{-2\Omega(s_3,s_2)}e^{-\Omega(s_2,T_l)}\left(1-e^{-\frac{\Delta_l}{\lambda_l}}\right)
\end{aligned} \tag{H.7}$$

For $i < j = k < l$ ($t_3 = s_2; t_2 > s_2$) and $W = 0$, $W' = 1$:

$$= \frac{2s_2+s_3}{2s_2+s_3-D_3}\,q(i,j,k,l) \tag{H.8}$$

For $i = k = l < j$ and $t_3 = s_3; t_2 < s_2$ and $W = W'$:

$$\begin{aligned}
&= \frac{2s_2+s_3}{2s_2+s_3-D_3}\,q^G(i,j,k,l)+ \\
&\quad \delta(W)\frac{\lambda_d\left(1-e^{\frac{D_3}{\lambda_d}}\right)}{2s_2+s_3-D_3}e^{-3\Omega(0,s_3)}\frac{1}{2}\left(1-e^{-\frac{2(T_{l+1}-s_3)}{\lambda_l}}\right)
\end{aligned} \tag{H.9}$$

Here, $q^G(i,j,k,l)$ is the contribution of (G.7) to $q(i,j,k,l)$ where $i = k = l < j$.

For $i = k = l < j$ and $t_3 < s_3; t_2 = s_3$, $W = 0$, and $W' = 1$:

$$\begin{aligned}
&= \frac{2s_2+s_3}{2s_2+s_3-D_3}\,q^{G2}(i,j,k,l)+ \\
&\quad \frac{\lambda_d\left(1-e^{\frac{D_3}{\lambda_d}}\right)}{2s_2+s_3-D_3}\frac{2}{3}e^{-3\Omega(0,T_k)}\left(1-e^{-\frac{3(s_3-T_k)}{\lambda_k}}\right)
\end{aligned} \tag{H.10}$$

Here, $q^{G2}(i,j,k,l)$ is the contribution of (G.8) to $q(i,j,k,l)$ where $i = k = l < j$.

For $t_3 < s_3; t_2 = s_3$, $W = 1$, $W' \in \{0, 1\}$:

$$= \frac{1}{2}\frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q^{G2}(i, j, k, l) \tag{H.11}$$

For $i < k = l = j$ $(t_3 > s_3\,; t_2 = s_2)$ and $W = W'$:

$$\begin{aligned} &= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q^{H}(i, j, k, l) + \\ &\qquad \delta(W-1)\frac{\lambda_d\left(1 - e^{\frac{D_3}{\lambda_d}}\right)}{2s_2 + s_3 - D_3} e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,T_k)}\left(1 - e^{-\frac{2(s_2 - T_k)}{\lambda_k}}\right) \end{aligned} \tag{H.12}$$

Here, $q^H(i, j, k, l)$ is the contribution of (G.9) to $q(i, j, k, l)$ where $i < k = l = j$.

For $i < k = l = j$ and $t_3 = s_2\,; t_2 > s_2$, $W = 0$, $W' = 1$:

$$= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q^{H2}(i, j, k, l) \tag{H.13}$$

Here, $q^{H2}(i, j, k, l)$ is the contribution of (G.10) to $q(i, j, k, l)$ where $i < k = l = j$.

For $i < k = l = j$ and $t_3 = s_2\,; t_2 > s_2$, $W = 1$, $W' \in \{0, 1\}$:

$$\begin{aligned} &= \frac{1}{2}\frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q^{H2}(i, j, k, l) + \\ &\qquad \delta(W')\frac{\lambda_d\left(1 - e^{\frac{D_3}{\lambda_d}}\right)}{2s_2 + s_3 - D_3} e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,s_2)}\left(1 - [1 - \delta(l-n)]e^{-\frac{T_{l+1} - s_2}{\lambda_l}}\right) \end{aligned} \tag{H.14}$$

For $i = j = k < l$ and $t_3 = s_3; t_2 > s_2$, $W' = W$:

$$\begin{aligned} &= \frac{2s_2 + s_3}{2s_2 + s_3 - D_3}\, q^{(I)}(i, j, k, l) + \\ &\qquad \delta(W)\frac{\lambda_d\left(1 - e^{\frac{D_3}{\lambda_d}}\right)}{2s_2 + s_3 - D_3} e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,s_2)} e^{-\Omega(s_2,T_l)}\left(1 - e^{-\frac{\Delta_l}{\lambda_l}}\right) \end{aligned} \tag{H.15}$$

Here, $q^{I}(i,j,k,l)$ is the contribution of (G.11) to $q(i,j,k,l)$ where $i=j=k<l$.

For $i=j=k<l$ and $t_3=s_2; t_2>s_2$, $W=0$, $W'=1$:

$$= \frac{2s_2+s_3}{2s_2+s_3-D_3}\, q^{I2}(i,j,k,l) \tag{H.16}$$

Here, $q^{I2}(i,j,k,l)$ is the contribution of (G.12) to $q(i,j,k,l)$ where $i=j=k<l$.

For $i=j=k<l$ and $t_3=s_2; t_2>s_2$, $W=1$, $W'\in\{0,1\}$:

$$\begin{aligned} &= \frac{1}{2}\frac{2s_2+s_3}{2s_2+s_3-D_3} q^{I2}(i,j,k,l) + \\ &\quad \delta(W')\frac{\lambda_d\left(1-e^{\frac{D_3}{\lambda_d}}\right)}{2s_2+s_3-D_3} e^{-3\Omega(0,s_3)} e^{-2\Omega(s_3,s_2)} e^{-\Omega(s_2,T_l)}\left(1-e^{-\frac{\Delta_l}{\lambda_l}}\right) \end{aligned} \tag{H.17}$$

For $k<i=j=l$ and $t_3<s_3; t_2=s_2$, $W'=W$:

$$= \frac{2s_2+s_3}{2s_2+s_3-D_3} q^{J}(i,j,k,l) + \delta(W-1)\frac{\lambda_d\left(1-e^{\frac{D_3}{\lambda_d}}\right)}{2s_2+s_3-D_3}\frac{2}{3} e^{-3\Omega(0,T_k)}\left(1-e^{-\frac{3\Delta_k}{\lambda_k}}\right) \tag{H.18}$$

Here, $q^{J}(i,j,k,l)$ is the contribution of (G.13) to $q(i,j,k,l)$ where $k<i=j=l$.

For $k<i=j=l$ and $t_3<s_3; t_2=s_3$, $W=0$, $W'=1$:

$$= \frac{2s_2+s_3}{2s_2+s_3-D_3} q^{J2}(i,j,k,l) + \frac{\lambda_d\left(1-e^{\frac{D_3}{\lambda_d}}\right)}{2s_2+s_3-D_3}\frac{2}{3} e^{-3\Omega(0,T_k)}\left(1-e^{-\frac{3\Delta_k}{\lambda_k}}\right) \tag{H.19}$$

Here, $q^{J2}(i,j,k,l)$ is the contribution of (G.14) to $q(i,j,k,l)$ where $k<i=j=l$.

For $k<i=j=l$ and $t_3<s_3; t_2=s_3$, $W=1$, $W'\in\{0,1\}$:

$$= \frac{1}{2}\frac{2s_2+s_3}{2s_2+s_3-D_3} q^{J2}(i,j,k,l) \tag{H.20}$$

References

[1] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson. Big data: Astronomical or genomical? *PLOS Biology*, 13(7):e1002195, 2015.

[2] J. Haldane. A mathematical theory of natural and artificial selection. *Transactions of the Cambridge Philosophical Society*, 23:19–41.

[3] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

[4] R. Fisher. *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford, 1930.

[5] M. Kimura. Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences*, 41:144–150, 1955.

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

[7] J. Wakeley, L. King, B. S. Low, and S. Ramachandran. Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman's Coalescent. *Genetics*, 190(4): 1433–1445, 2012.

[8] S. Carmi, P. R. Wilton, J. Wakeley, and I. Pe'er. A renewal theory approach to IBD sharing. *Theoretical Population Biology*, 97:35–48, 2014.

[9] P. F. Palamara, L. C. Francioli, P. R. Wilton, G. Genovese, A. Gusev, H. K. Finucane, S. Sankararaman, S. R. Sunyaev, P. I. W. de Bakker, J. Wakeley, I. Pe'er, and A. L. Price. Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. *The American Journal of Human Genetics*, 97(6):775–789, 2015.

[10] G. A. T. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

[11] P. Marjoram and J. D. Wall. Fast "coalescent" simulation. *BMC Genetics*, 7(1): 16, 2006.

[12] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

[13] S. Sheehan, K. Harris, and Y. S. Song. Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2013.

[14] M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. *PLOS Genetics*, 10(5):e1004342, 2014.

[15] P. R. Wilton, S. Carmi, and A. Hobolth. The SMC' is a highly accurate approximation to the ancestral recombination graph. *Genetics*, 200(1):343–355, 2015.

[16] A. Hobolth, O. F. Christensen, T. Mailund, and M. H. Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLOS Genetics*, 3(2):e7, 2007.

[17] S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, 2014.

[18] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

[19] J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press, Oxford, 1 edition, 2005.

[20] J. Wakeley. *Coalescent Theory: An Introduction.* Roberts and Co., Greenwood Village, CO, 2009.

[21] J. M. Pemberton. Wild pedigrees: The way forward. *Proceedings of the Royal Society B: Biological Sciences*, 275:613–621, 2008.

[22] M. Larmuseau, A. Van Geystelen, M. van Oven, and R. Decorte. Genetic genealogy comes of age: Perspectives on the use of deep-rooted pedigrees in human population genetics. *American Journal of Physical Anthropology*, 150(4):505–511, 2013.

[23] J. T. Chang. Recent common ancestors of all present-day individuals. *Advances in Applied Probability*, 31(4):1002–1026, 1999.

[24] B. Derrida, S. C. Manrubia, and D. H. Zenette. On the genealogy of a population of biparental individuals. *Journal of Theoretical Biology*, 203(3):303–315, 2000.

[25] D. L. T. Rohde, S. Olson, and J. T. Chang. Modelling the recent common ancestry of all living humans. *Nature*, 431:562–566, 2004.

[26] N. H. Barton and A. M. Etheridge. The relation between reproductive value and genetic contribution. *Genetics*, 188(4):953–973, 2011.

[27] R. R. Hudson. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[28] S. Wright. The genetical structure of populations. *Annals of Eugenics*, 15(4): 323–354, 1951.

[29] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[30] R. C. Griffiths and S. Tavaré. Ancestral Inference in Population Genetics. *Statistical Science*, 9(3):307–319, 1994.

[31] Y. Wu. Exact computation of coalescent likelihood for panmictic and subdivided populations under the infinite sites model. *IEEE Transactions on Computational Biology and Bioinformatics*, 7:611–618, 2010.

[32] A. Raj, M. Stephens, and J. K. Pritchard. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589, 2014.

[33] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

[34] H. Tang, J. Peng, P. Wang, and N. J. Risch. Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, 2005.

[35] I. Moltke and A. Albrechtsen. RelateAdmix: A software tool for estimating relatedness between admixed individuals. *Bioinformatics*, 30(7):1027–1028, 2014.

[36] T. Thornton, H. Tang, T. Hoffmann, H. Ochs-Balcom, B. Caan, and N. Risch. Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1):122–138, 2012.

[37] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[38] G. A. Wilson and B. Rannala. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, 163(3):1177–1191, 2003.

[39] A. Bhaskar, A. G. Clark, and Y. S. Song. Distortion of genealogical properties when the sample is very large. *Proceedings of the National Academy of Sciences*, 111(6):2385–2390, 2014.

[40] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genetics*, 5(10):e1000695, 2009.

[41] J. A. Kamm, J. Terhorst, and Y. S. Song. Efficient computation of the joint sample frequency spectra for multiple populations. *Journal of Computational and Graphical Statistics*, pages 1–37, 2016.

[42] S. Gazal, M. Sahbatou, M.-C. Babron, E. Génin, and A.-L. Leutenegger. High level of inbreeding in final phase of 1000 Genomes Project. *Scientific Reports*, 5, 2015.

[43] T. J. Pemberton, C. Wang, J. Z. Li, and N. A. Rosenberg. Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *The American Journal of Human Genetics*, 87(4):457–464, 2010.

[44] N. A. Rosenberg. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*, 70(6):841–847, 2006.

[45] E. A. Thompson. Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, 2013.

[46] J. E. Powell, P. M. Visscher, and M. E. Goddard. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, 11:800–805, 2010.

[47] P. Ralph and G. Coop. The geography of recent genetic ancestry across europe. *PLOS Biology*, 11(5):e1001555, 2013.

[48] S. R. Browning and B. L. Browning. Identity by descent between distant relatives: Detection and applications. *Annual Review of Genetics*, 46(1):617–633, 2012.

[49] P. F. Palamara, T. Lencz, A. Darvasi, and I. Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics*, 91(5):809–822, 2012. ISSN 0002-9297.

[50] P. Tataru, J. A. Nirody, and Y. S. Song. diCal-IBD: Demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics*, 30(23): 3430–3431, 2014.

[51] B. Browning and S. Browning. Detecting identity by descent and estimating genotype error rates in sequence data. *The American Journal of Human Genetics*, 93(5):840–851, 2013. ISSN 0002-9297.

[52] K. Harris and R. Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLOS Genetics*, 9(6):e1003521, 2013.

[53] R. Hudson. Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics, editors, *Oxford Surveys in Evolutionary Biology*, volume 7, pages 1–44. 1991.

[54] R. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution*, volume 87 of *IMA Volumes in Mathematics and Its Application*, pages 257–270. Springer-Verlag, New York, 1997.

[55] C. Wiuf and J. Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259, 1999.

[56] J. Y. Dutheil, G. Ganapathy, A. Hobolth, T. Mailund, M. K. Uyenoyama, and M. H. Schierup. Ancestral population genomics: The coalescent hidden Markov model approach. *Genetics*, 183(1):259–274, 2009.

[57] N. Kaplan and R. R. Hudson. The use of sample genealogies for studying a selectively neutral m-loci model with recombination. *Theoretical Population Biology*, 28(3):382–396, 1985.

[58] K. L. Simonsen and G. A. Churchill. A Markov chain model of coalescence with recombination. *Theoretical Population Biology*, 52(1):43–59, 1997.

[59] A. Hobolth and J. L. Jensen. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical Population Biology*, 2014.

[60] E. A. van Doorn and A. I. Zeifman. Birth-death processes with killing. *Statistics & Probability Letters*, 72(1):33–42, 2005.

[61] A. Eriksson, B. Mahjani, and B. Mehlig. Sequential Markov coalescent algorithms for population models with demographic structure. *Theoretical Population Biology*, 76(2):84–91, 2009.

[62] R. C. Griffiths. Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology*, 19(2):169–186, 1981.

[63] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of dna sequences. *Genetics*, 111(1): 147–164, 1985.

[64] T. Ohta and M. Kimura. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics*, 68(4):571–580, 1971.

[65] G. A. T. McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162(2):987–991, 2002.

[66] G. Schwarz. Noninvariance of $\bar{d}$-convergence of k-step Markov approximations. *Annals of Probability*, 4(6):1033–1035, 1976.

[67] R. Fernández and A. Galves. Markov approximations of chains of infinite order. *Bulletin of the Brazilian Mathematical Society*, 33(3):1–12, 2002.

[68] S. Gallo, M. Lerasle, and D. Takahashi. Markov approximation of chains of infinite order in the $\bar{d}$-metric. *Markov Processes and Related Fields*, 19(1):51–82, 2013.

[69] J. Kim, E. Mossel, M. Z. Rácz, and N. Ross. Can one hear the shape of a population history? *arXiv:1402.2424*, 2014.

[70] C. Wiuf. Consistency of estimators of population scaled parameters using composite likelihood. *Journal of Mathematical Biology*, 53(5):821–841, 2006.

[71] G. K. Chen, P. Marjoram, and J. D. Wall. Fast and flexible simulation of DNA sequence data. *Genome Research*, 19(1):136–142, 2009.

[72] J. Lehtonen, M. D. Jennions, and H. Kokko. The many costs of sex. *Trends in Ecology & Evolution*, 27(3):172–178, 2012.

[73] S. Meirmans and R. Strand. Why are there so many theories for sex, and what do we do with them? *Journal of Heredity*, 101:S3–S12, 2010.

[74] M. J. McDonald, D. P. Rice, and M. M. Desai. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature*, 531(7593):233–236, 2016.

[75] L. Becks and A. F. Agrawal. The evolution of sex is favoured during adaptation to new environments. *PLOS Biology*, 10(5):e1001317, 2012.

[76] C. M. Lively and M. F. Dybdahl. Parasite adaptation to locally common host genotypes. *Nature*, 405(6787):679–681, 2000.

[77] J. Bast, I. Schaefer, T. Schwander, M. Maraun, S. Scheu, and K. Kraaijeveld. No accumulation of transposable elements in asexual arthropods. *Molecular Biology and Evolution*, 33(3):697–706, 2016.

[78] M. Neiman, J. Jokela, and C. M. Lively. Variation in asexual lineage age in *Potamopyrgus antipodarum*, a New Zealand snail. *Evolution*, 59(9):1945–1952, 2005.

[79] J. Maynard Smith. *The Evolution of Sex.* Cambridge Univ. Press, London, 1978.

[80] J. A. Palacios, J. Wakeley, and S. Ramachandran. Bayesian nonparametric inference of population size changes from sequential genealogies. *Genetics*, 201(1): 281–304, 2015.

[81] K. Harris, S. Sheehan, J. A. Kamm, and Y. S. Song. Decoding coalescent hidden Markov models in linear time. In R. Sharan, editor, *Research in Computational Molecular Biology*, number 8394 in Lecture Notes in Computer Science, pages 100–114. Springer International Publishing, 2014.

[82] A. Ceplitis. Coalescence times and the Meselson effect in asexual eukaryotes. *Genetics Research*, 82(03):183–190, 2003. ISSN 1469-5073.

[83] J. Kelleher, A. M. Etheridge, and G. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *bioRxiv*, page 033118, 2015.

[84] C. Wiuf and J. Hein. The coalescent with gene conversion. *Genetics*, 155(1): 451–462, 2000.

[85] J.-F. Flot, B. Hespeels, X. Li, B. Noel, I. Arkhipova, E. G. J. Danchin, A. Hejnol, B. Henrissat, R. Koszul, J.-M. Aury, V. Barbe, R.-M. Barthélémy, J. Bast, G. A. Bazykin, O. Chabrol, A. Couloux, M. Da Rocha, C. Da Silva, E. Gladyshev, P. Gouret, O. Hallatschek, B. Hecox-Lea, K. Labadie, B. Lejeune, O. Piskurek, J. Poulain, F. Rodriguez, J. F. Ryan, O. A. Vakhrusheva, E. Wajnberg, B. Wirth, I. Yushenova, M. Kellis, A. S. Kondrashov, D. B. Mark Welch, P. Pontarotti, J. Weissenbach, P. Wincker, O. Jaillon, and K. Van Doninck. Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga. Nature*, 500:453–457, 2013.

[86] M. Hartfield, S. I. Wright, and A. F. Agrawal. Coalescent times and patterns of genetic diversity in species with facultative sex: Effects of gene conversion, population structure, and heterogeneity. *Genetics*, 202(1):297–312, 2016.

[87] J. Yin, M. I. Jordan, and Y. S. Song. Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. *Bioinformatics*, 25 (12):i231–i239, 2009.

[88] J. Turgeon and P. D. N. Hebert. Evolutionary interactions between sexual and all-female taxa of *Cyprinotus* (Ostracoda: Cyprididae). *Evolution*, 48(6):1855–1865, 1994.

[89] Z. Majtánová, L. Choleva, R. Symonová, P. Ráb, J. Kotusz, L. Pekárik, and K. Janko. Asexual reproduction does not apparently increase the rate of chromosomal evolution: Karyotype stability in diploid and triploid clonal hybrid fish (*Cobitis*, Cypriniformes, Teleostei). *PLOS ONE*, 11(1):e0146872, 2016.

[90] S. B. J. Menken, E. Smit, and H. J. C. M. D. Nijs. Genetical population structure in plants: Gene flow between diploid sexual and triploid asexual dandelions (*Taraxacum* section *Ruderalia*). *Evolution*, 49(6):1108–1118, 1995.

[91] R. D. Noyes and L. H. Rieseberg. Two independent loci control agamospermy (apomixis) in the triploid flowering plant *Erigeron annuus*. *Genetics*, 155(1): 379–390, 2000.

[92] E. Y. Y. Lo, S. Stefanović, and T. A. Dickinson. Population genetic structure of diploid sexual and polyploid apomictic hawthorns (*crataegus*; Rosaceae) in the Pacific Northwest. *Molecular Ecology*, 18(6):1145–1160, 2009.

[93] G. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, 1975.

[94] K. Itō. *Essentials of Stochastic Processes*. American Mathematical Society, 2006.

[95] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, 1975.

Supplementary Material

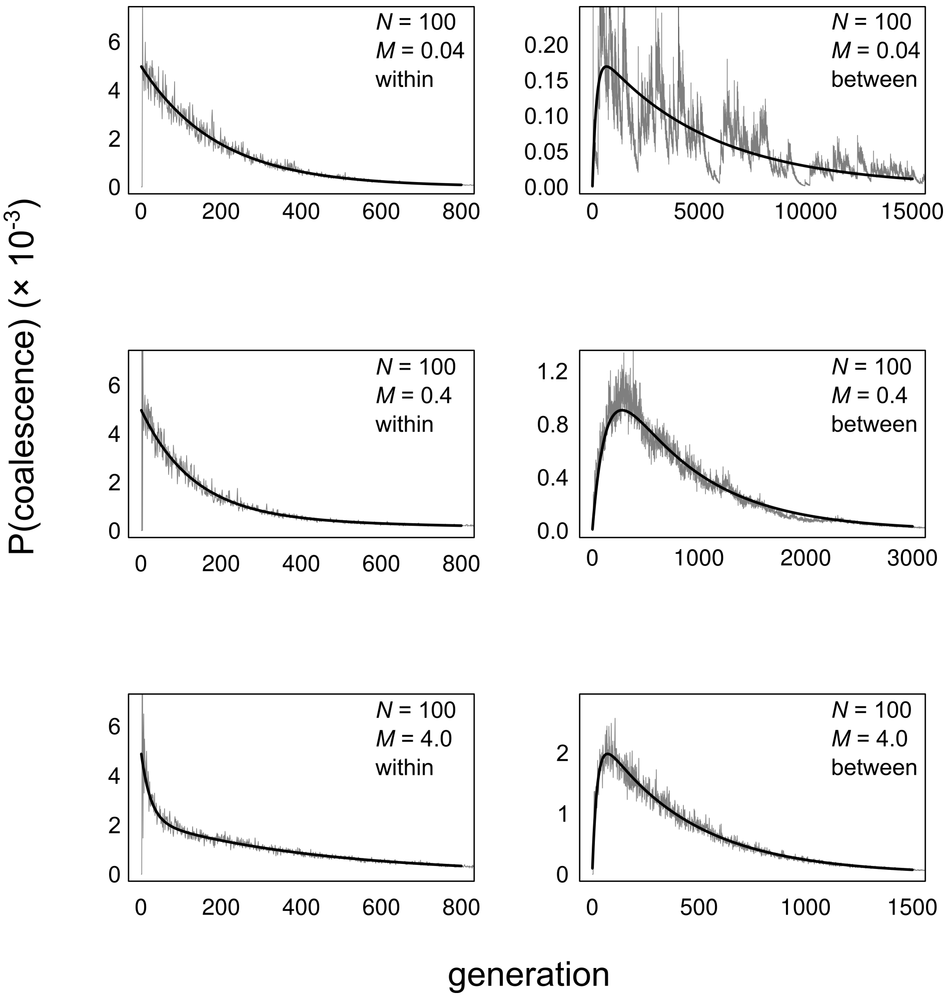

Figure S1: Distributions of coalescence times in two-deme populations with $N = 100$ individuals in each deme. Each panel shows a distribution of coalescence times for a particular value of the migration rate $M = 4Nm$. For panels in the left column, two individuals were sampled from the same deme ("within-deme" sampling), and in the right column two individuals were sampled from different demes.

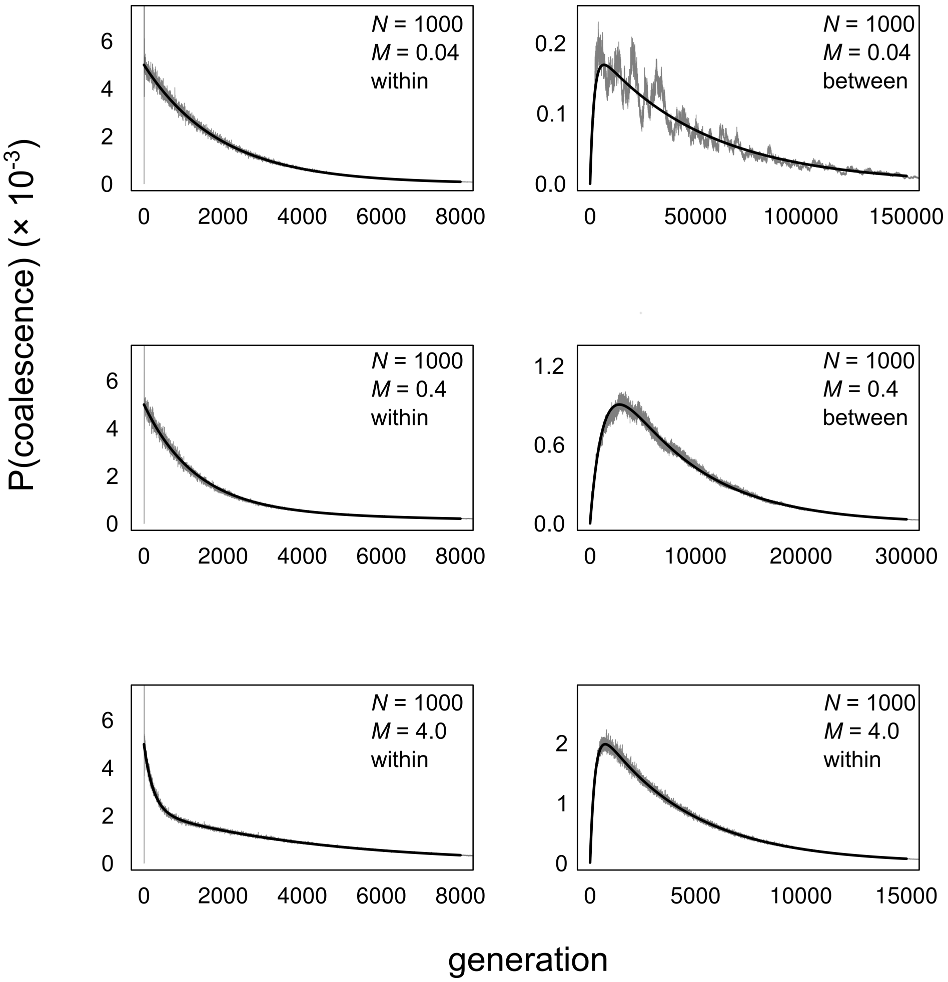

Figure S2: Distributions of coalescence times in two-deme populations with $N = 1000$ individuals in each deme. Each panel shows a distribution of coalescence times for a particular value of the migration rate $M = 4Nm$. For panels in the left column, two individuals were sampled from the same deme ("within-deme" sampling), and in the right column two individuals were sampled from different demes.

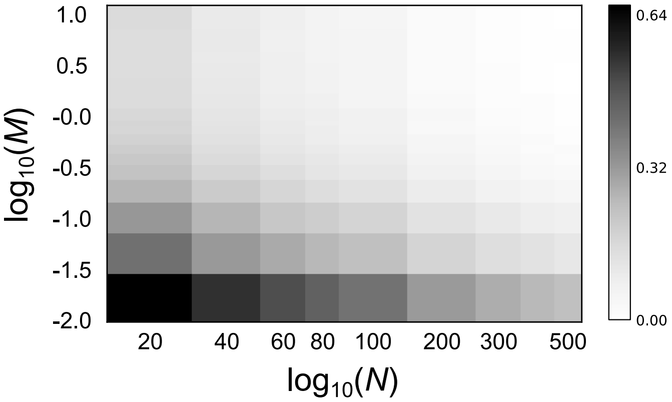

Figure S3: Total variation distance between pairwise coalescence time distributions in pedigrees versus standard theory. For each point on the grid, the total variation distance from the prediction of standard theory was averaged over 20 pedigrees with the corresponding N and M.

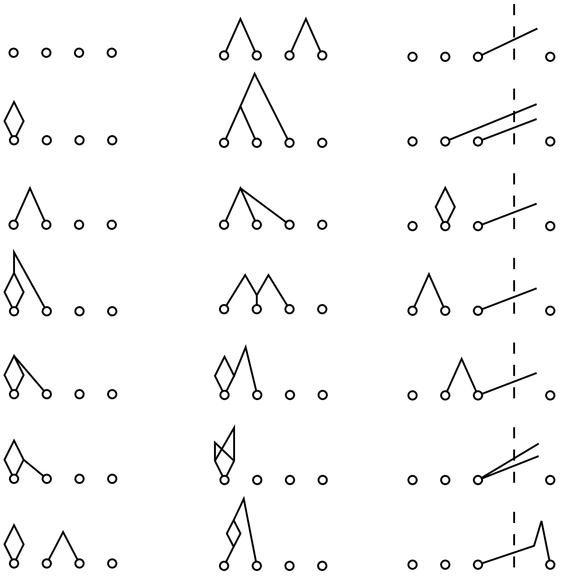

Figure S4: Distinct Wright-Fisher pedigree shapes with two or fewer IBD or admixture events in a two-deme population. Only relationships involved in IBD or admixture events are shown. All pedigrees have non-overlapping generations, and sampled individuals (white circles) are living in the present generation. To produce all distinct pedigrees, it is necessary to consider all the ways of indexing the individuals involved in the IBD and admixture events, as well as the all of the possible timings of these events.

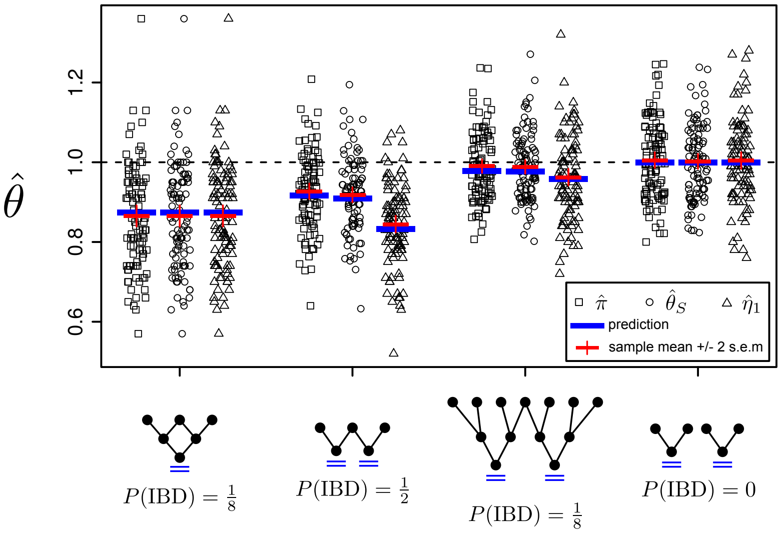

Figure S5: Estimates of $\theta = 4N\mu$ based on 100 replicate simulations of 200 loci segregating through sample pedigrees featuring identity by descent. Blue lines indicate theoretical predictions for individual pedigrees and estimators. Red horizontal lines indicate sample means, and red vertical lines indicate twice the standard error of the mean. The true value of $\theta = 1.0$ is indicated by the dashed line. Note that for the first pedigree, with $n = 2$, the three estimators are equivalent.

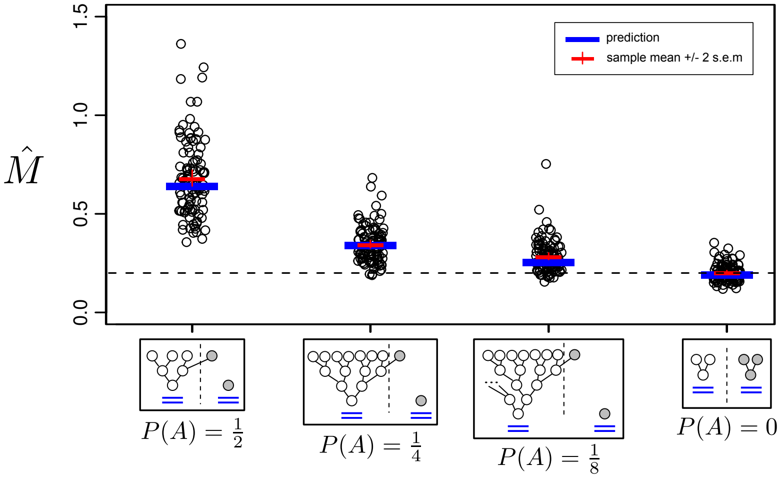

Figure S6: Estimates of M from the estimator given by $(\text{B}.3)$ in replicate simulated genetic datasets of 100 loci segregating through sample pedigrees containing recent admixture. Blue lines indicate theoretical predictions for individual pedigrees. Red horizontal lines indicate sample means, and red vertical lines indicate twice the standard error of the mean. The true value of $M = 0.2$ is indicated by the dashed line. $P(A)$ indicates the probability of admixture in the indicated pedigrees.

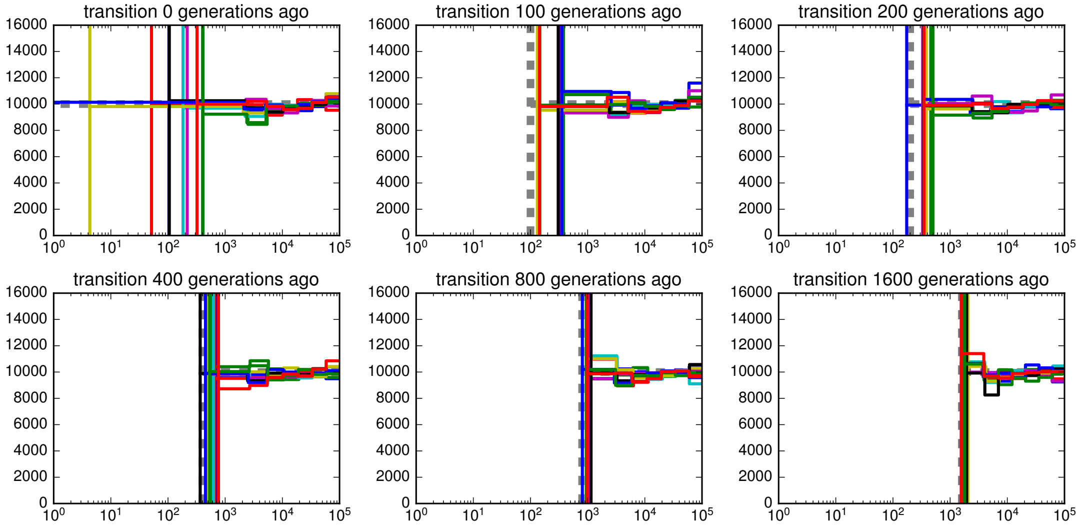

Figure S7: Inferred demographic history of asexual lineages of different ages with constant-sized ancestral sexual populations. Each panel shows the inferred sexual-to-asexual transition time and population size history of the sexual ancestor for five replicate simulations. The true values are shown with a thick gray dashed line, and the inferred history of each replicate simulation is shown with a differently colored thin solid line. Vertical lines show the inferred transition time, and horizontal lines to the right of the transition show the inferred population size history of the sexual ancestor. In each simulation, the size of the sexual ancestral population was $N = 10000$, the mutation rate was 1.5×10^{-8} per generation per base pair, and the recombination rate was 1.0×10^{-8} per generation per base pair. Each simulated genome was 120 Mbp in length.

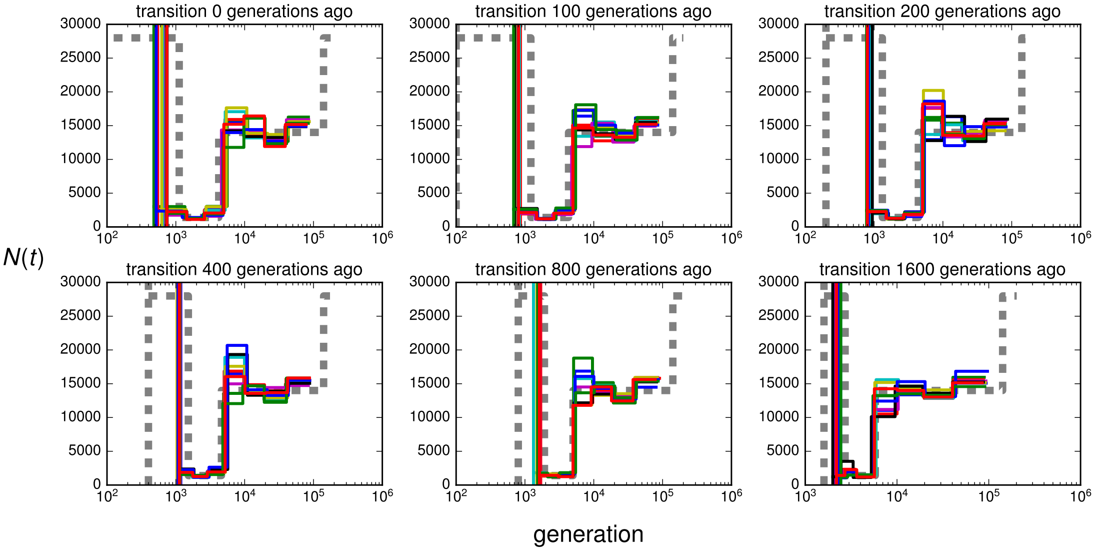

Figure S8: Inferred asexual lineage age and ancestral sexual population history for recently derived asexual triploid lineages whose sexual ancestors underwent a bottleneck just prior to the onset of asexual reproduction. Each panel shows the inferred sexual-to-asexual transition time and population size history of the sexual ancestor for five replicate simulations. The true values are shown with a thick gray dashed line, and the inferred history of each replicate simulation is shown with a differently colored thin solid line. Vertical lines show the inferred transition time, and horizontal lines to the right of the transition show the inferred population size history of the sexual ancestor. In each simulation, the size of the sexual ancestral population was $N = 10000$, the mutation rate was 1.5×10^{-8} per generation per base pair, and the recombination rate was 1.0×10^{-8} per generation per base pair. Each simulated genome was 120 Mbp in length.

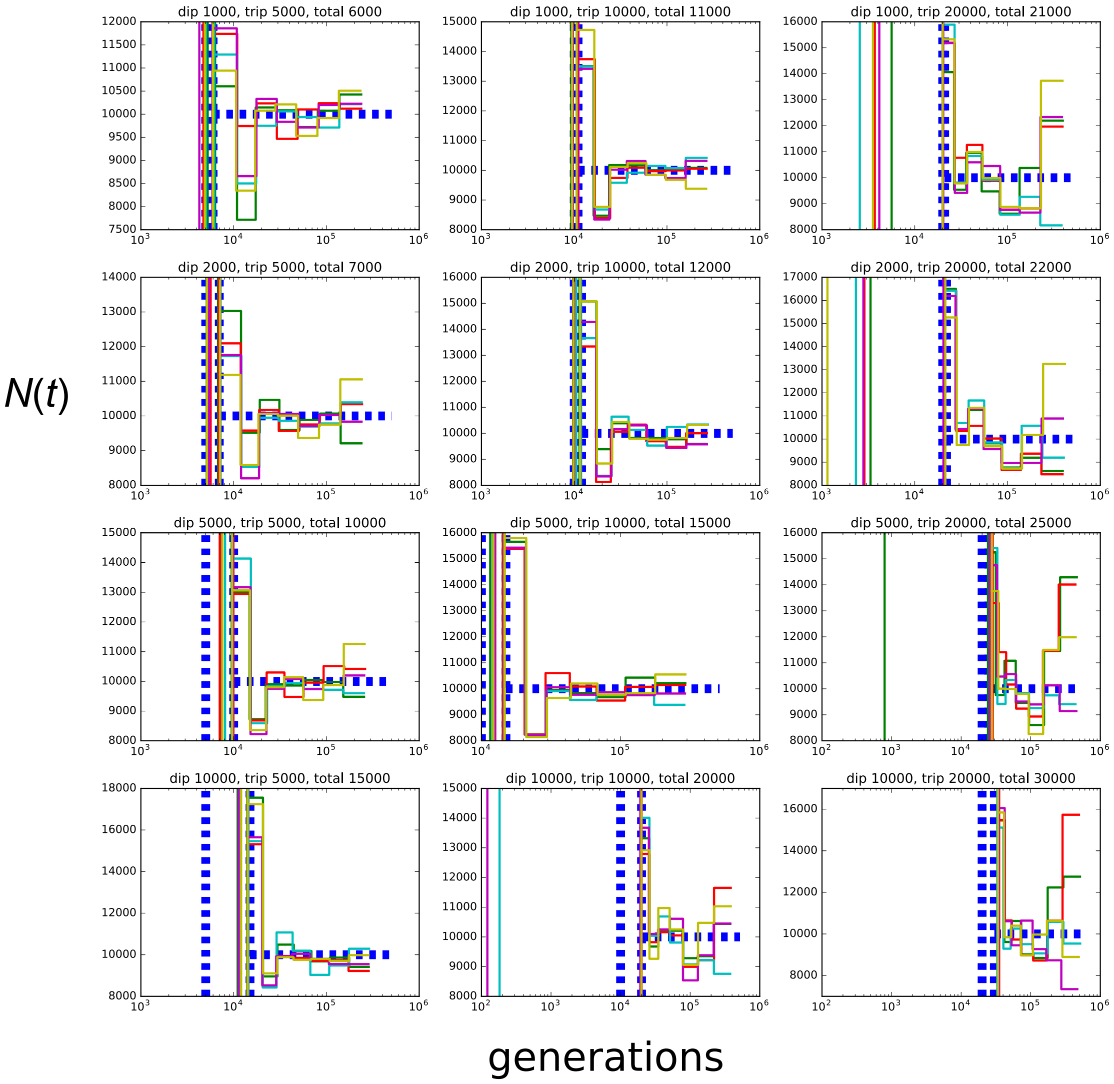

Figure S9: Inferred lengths of diploid and triploid asexual reproduction intervals for triploid asexual lineages that were formed when a diploid asexual lineage incorporated a haploid, sexual sperm. Each panel shows inference made from five replicate simulations of asexual lineages with a particular length of triploid and diploid asexual reproduction. The two vertical dashed lines show the true diploid and triploid transition times (triploid to the left, diploid to the right), measured from the present. The two vertical lines of a particular color show the inferred starts of diploid and triploid reproduction, and the horizontal colored lines show the inferred population size history of the diploid asexual ancestor. The true population size is $N = 10000$, the mutation rate is $\mu = 1.5 \times 10^{-8}$, the recombination rate is $r = 1.0 \times 10^{-8}$, and 100 Mbp was simulated for each replicate genome.