THE UNIVERSITY OF CHICAGO

A NOVEL POPULATION GENOMIC METHOD FOR INFERRING POPULATION HISTORY AND DETECTING ADAPTIVE GENETIC VARIATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHYSICS

BY
GAUTAM UPADHYA

CHICAGO, ILLINOIS
DECEMBER 2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Foremost, I’d like to thank Matthias Steinrücken, my advisor, for his mentorship and support with these projects. I wouldn’t have been able to complete my PhD without his guiding hand, and I’m truly grateful that our paths have crossed. I’d also like to thank the members of our lab, Maryn Carlson, Eric Friedlander, and Xiaoheng Cheng. Our friendships and discussions have been invaluable.

Next, I’d like to thank John Novembre, Jeremy Berg, their labs, and the larger popgen community at UChicago. Our group meetings provided me a fertile environment to generate new ideas, modify existing approaches, and simply learn much more. In particular, Arjun Biddanda and Dan Rice have provided valuable feedback and thought provoking suggestions, and Margarita Orlova assisted with setting up the initial simulation studies to aid this work.

Finally, I’d like to thank my friends and family from the bottom of my heart: Mom, dad, Vijay, Brie, and the many, many more who have encouraged and believed in me. To Nikita, I don’t think I can express my full gratitude, so an ”I love you,” will have to suffice.

ABSTRACT

Advances in genetic sequencing technology have given us an abundance of whole genome sequencing data that can tell us much about evolutionary processes and population history. Coalescent hidden Markov models (CHMMs) are a powerful class of methods that use both genetic variation and genetic linkage information to untangle the complex demographic histories of natural populations. This thesis presents `CHIMP` (CHMM History-Inference ML Procedure), a novel CHMM implementation that can use both the height (T_{MRCA}) and the total branch length ($\mathcal{L}$) of the underlying genealogical tree as the latent variable in the HMM as the method moves sequentially along the genome. The primary application of `CHIMP` is in demographic inference problems, and we perform a suite of simulations to benchmark the performance of `CHIMP` among other state of the art CHMMs. We also demonstrate the use of `CHIMP` to perform demographic inference in structured populations. Finally we introduce `CHIMP-PD`, an extension that is used to decode the posterior probability of the CHMM, and explore its use in uncovering patterns of adaptive variation. This work ultimately demonstrates that `CHIMP` provides a flexible, efficient alternative to other methods, particularly when analyzing unphased and pseudohaploid data.

CHAPTER 1

INTRODUCTION

Technological advances in full-genome sequencing in the recent decades have made it possible to collect increasingly large amounts of genomic data sampled from individuals across diverse population groups and even species. These data can be used to study relatedness among individuals and can also infer complex demographic histories, helping to unravel population size changes, population structure and migration events. In addition, adaptation of beneficial alleles and other forms of genetic selection often leave characteristic signatures in the genome that can be studied to better understand the selective processes that continue to shape our DNA. Thus, this wealth of genomic data gives us a unique window into many of the forces that shape our genetic material.

Population history is one such force that has been of particular interest to population geneticists. Understanding the evolutionary history of humans can be both culturally and anthropologically illuminating, but it is also important in other medical and health-related applications. For example, genome wide association studies (GWAS), a popular tool used to detect associations between genetic factors and specific phenotypic outcomes, can be confounded by genetic variation arising from demographic history [1]. Thus it is crucial to develop population genetic tools for analyzing whole-genome sequencing data that can infer the underlying demographic history and establish appropriate null models for these associations and other studies.

To date, many methods have been presented in the literature that infer different aspects of demographic histories from different signals in the data. The primary focus in this study is on the inference of the size history of a single population, and here we briefly review methods with a similar focus. Several methods perform inference using the site frequency spectrum (SFS), either assuming no linkage between sites [2, 3], or complete linkage [4]. These methods can be efficiently applied to large sample sizes which particularly improves their ability to infer recent population size changes [3, 5]. However, these methods do not leverage information about decay of linkage disequilibrium along the chromosome, which has been shown to increase power. Other methods make use of linkage information by fitting demographic models to the empirical distribution of long shared tracts of Identity-By-Descent directly [6, 7]. Since most such methods consider tracts above a certain length threshold, they are also most powerful at inferring recent population size changes. While these methods account for some linkage information, they do not model the correlation in tract length along the genome. Some recent methods aim to directly reconstruct the multi-locus genealogy relating the sampled individuals from high-quality genomic sequencing data [8, 9, 10]. Such genealogies are useful for a variety of down-stream analyses and can be used for demographic inference as well.

A powerful class of methods to infer population size histories that account for linkage, both in terms of length of shared haplotypes and correlation along the genome, are Coalescent Hidden Markov Models (CHMMs). These methods are based on the Sequentially Markovian Coalescent (SMC) introduced by [11], based on work by [12]. In this framework, the correlations among the marginal genealogies relating the sampled individuals at each locus in the genome due to chromosomal linkage and ancestral recombination events is approximated by a Markov chain. The observed genetic variation is subsequently modeled by a mutation process on these marginal genealogies. Using the full marginal genealogies as latent states in a Hidden Markov Model (HMM) framework is prohibitive, but employing lower-dimensional summaries of these genealogies facilitates computationally efficient inference of population size histories.

A number of different CHMM-based inference tools have been developed, including `PSMC` [13], `MSMC` [14], `MSMC2` [15], `SMC++` [16], and `diCal` [17, 18]. These methods differ in the sample size that they can analyze and in how the marginal genealogies are represented in the respective CHMM. For example, `PSMC` can only be applied to samples of size 2, whereas

MSMC2 can process more samples. However, the computational cost of the latter does increase substantially with sample size. SMC++ can be applied to large samples and the data does not need to be phased, whereas diCal requires phased data and is only applicable to moderate sample sizes. The specific implementation details result in each method performing well for certain sample sizes and for certain time periods [19, 20], but no method performs uniformly well across all parameter regimes.

This thesis presents a novel CHMM method, CHIMP (**C**HMM **H**istory-**I**nference **ML** **P**rocedure), that can handle unphased and pseudo-haploid data and scales efficiently to large sample sizes. Chapter 2 describes the modeling framework for CHIMP and describes our two implementations that differ in the hidden state space that they use for the CHMM. One implementation uses the T_{MRCA}, the time to most recent common ancestor of the local genealogical tree, while the other uses $\mathcal{L}$, the total branch length of the tree. This chapter also describes how to numerically solve certain systems of differential equations to compute the necessary transition and emission probabilities for the CHMM, thus extending previous work done on the subject [21]. Chapter 3 incorporates the CHMM model presented in Chapter 2 into a standard Expectation-Maximization (EM) framework to perform demographic inference and presents the results of a suite of tests comparing CHIMP to other similar methods. Chapter 4 explores how CHIMP can be applied in cases with more structured populations. Finally, Chapter 5 uses the machinery of CHIMP to study the posterior distribution of the latent state. We demonstrate how tree shape can be affected by selection, and how CHIMP-PD can be used as a tool for detecting adaptive variation.

The methods and results presented in Chapters 2 and 3 are available as a pre-print on the bioRxiv and are under review for further publication. The software for CHIMP is available at https://github.com/steinrue/chimp.

CHAPTER 2

CHIMP - MODEL AND IMPLEMENTATION

In this chapter we will provide the relevant background on the Sequentially Markovian Coalescent (SMC) which is the basis for CHMMs. We will then introduce our novel CHMM model, `CHIMP`, including two versions that use different representations of the local genealogical trees as the hidden state: one uses the tree height T_{MRCA}, the other uses the total branch length $\mathcal{L}$. We will describe how we compute the necessary probabilities for our HMM for each of these versions, and will introduce the likelihood models and computational algorithms that we use. Finally in section 2.6, we will describe some details regarding our software implementation.

2.1 Background on CHMMs

The genetic variation observed in a sample of n haploid sequences from a given population is affected by its population size history $N(k)$, where $N(k)$ is the number of diploid individuals in the population k generations before present. We use coalescent theory to model the effects that a time varying population size has on the genealogy relating a sample together, which in turn affects the pattern of observed genetic variation. In the coalescent framework, it is convenient to measure time in units of $2N(0)$ generations and to consider the population size relative to the size at present. To this end, we introduce the relative coalescent-scaled population size

$$\eta(t) := \frac{N(2N_0t)}{N_0},$$

where $k = 2N_0t$ is the rescaled time and $N_0 \equiv N(0)$ is the present population size, but can just as easily be an arbitrarily chosen reference size.

The single-locus coalescent models the genetic variation among n sampled haploids at a particular locus in the genome [22]. In the coalescent, the genealogy of the sample is described by following the ancestral lineages of the n haplotypes (sampled at present) back in time. Each pair of lineages can coalesce (find a common ancestor) at a given rate $\lambda(t)$ that can vary with time t. The coalescent rate is the inverse of the relative population size $\lambda(t) = 1/\eta(t)$, which reflects the fact that ancestral lineages coalesce faster in small populations but coalesce more slowly in larger populations. This process proceeds until all lineages coalesce into a single lineage, referred to as the most recent common ancestor (MRCA), the genetic ancestor of all haplotypes in the sample. The time of this final coalescent event is denoted by T_{MRCA}. The coalescent thus gives the distribution of genealogies at a single locus. One can model the observed genetic variation at the given locus by superimposing mutations on the genealogy according to a Poisson process with rate $\theta/2$, where $\theta = 4N_0\mu$ is the population-scaled mutation parameter and μ is the per-generation per-site mutation probability.

The standard coalescent models the marginal genealogy at a single locus. To analyze genomic sequence data, one can use the ancestral recombination graph (ARG), which extends the regular coalescent model to describe the full multi-locus genealogy for n sampled haplotypes across L loci [23, 24]. Specifically, the ARG models the genealogies at each individual locus and their correlations induced by the presence or absence of ancestral recombination events. Just as in the single-locus case, mutations can be superimposed onto these genealogies to model the observed genetic variation in multi-locus genomic sequence data. While the ARG is a useful tool to simulate genomic data [24, 25], in many scenarios its applicability in likelihood-based population genetic inference is hindered by its complexity: The space of possible ARGs grows quickly with the number of samples and the length of the genome.

One factor contributing to the complexity of the ARG is the fact that the marginal genealogies at distant loci can depend on each other [12]. The Sequentially Markovian Coalescent (SMC), introduced by [11], simplifies the model by assuming that the distribution of the marginal genealogy at a given locus only depends on the genealogy at the previous locus in the sequence, that is, it assumes that the sequence of marginal genealogies is a Markov

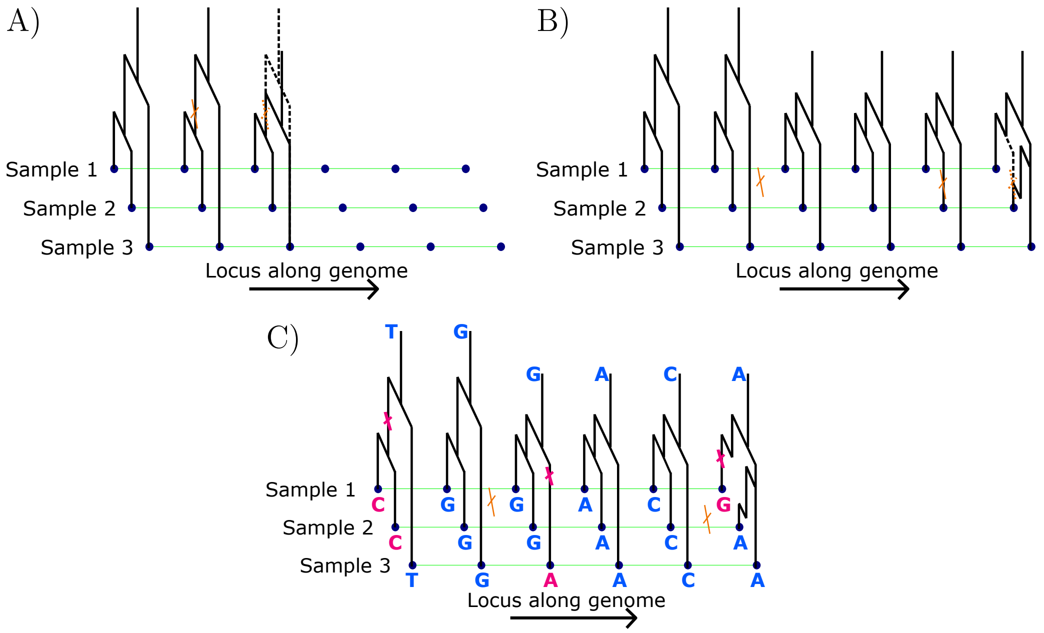


Figure 2.1: Panel A) shows the marginal genealogy being propagated unchanged along the genome until an ancestral recombination event is encountered, and the genealogy modified accordingly. In panel B), the new genealogy is propagated until a second recombination event is encountered. Panel C) demonstrates a realization of the mutation process along the genealogy at each locus and the resulting observed genetic data, where the ancestral allele is changed to a derived allele at samples that are subtended by the mutation event.

chain. Under the SMC, the sequence of marginal genealogies is generated as follows. The genealogy at the first locus is distributed according to the standard coalescent. To proceed from one locus to the next, ancestral recombination events occur according to a Poisson process at rate $\frac{\rho}{2}$ on the branches of the current genealogy, where $\rho = 4N_0r$, and r is the per base-pair per generation recombination probability. If no recombination events occur, the marginal genealogy is copied unchanged to the next locus. However, if recombination does occur, the lineage above the recombination event is removed up to the next coalescent event involving this lineage. To obtain the genealogy at the next locus, the removed lineage is then replaced by a new lineage that undergoes the standard coalescent dynamic, that is, it can coalesce with the regular coalescent rate into the other ancestral lineages. The distribution of the genealogical trees at each locus is fully determined by the genealogy at the previous locus in the sequence, and thus the sequence of genealogical trees is a Markov chain. An illustration of this generative process for the marginal genealogies is depicted in panels A) and B) of Figure 2.1.

CHMMs use the SMC as the basis for computing likelihoods of observed genomic sequence data. Given the local genealogy at a locus, we can imagine that ancestral mutations may have occurred along the genealogical branches. By modeling mutations as a Poisson process on these trees we can produce mutations that are inherited by our present-day samples, with patterns of variation that are dependent on the tree structure. In this way, the likelihood of the observed genetic variation can be computed as emissions conditional on the hidden state at a given locus (panel C of Figure 2.1).

2.2 CHIMP: A CHMM Using T_{MRCA} or $\mathcal{L}$

While it is possible, in principle, to implement the full SMC as a CHMM by representing the full local genealogy as a hidden state (as depiceted in Figure 2.1), such an implementation would make likelihood-based inference intractable due to the prohibitively large hidden state space resulting from the continuous nature of the genealogical times and the fact that the number of topologies grows super-exponentially with the number of samples. Thus, most existing implementations of CHMMs use a suitable discretization of time and approximate the full local genealogical trees using lower-dimensional summaries (often only one-dimensional) to arrive at a finite hidden states space for the HMM (a summary of some common CHMM implementations and their choices for the hidden states can be found in [19, Fig 1]). Our method, `CHIMP`, is a CHMM with a one-dimensional hidden state. We either use the T_{MRCA} (time to most recent common ancestor, i.e. tree height) or $\mathcal{L}$ (total branch length of the tree) as the hidden state. We use $S+1$ increasing times (or lengths) $t_0 = 0 < t_1 < \ldots < t_S = \infty$ to partition the positive real numbers into S discrete intervals. The CHMM is in state s_i at locus ℓ if $t_{i-1} \leq T_\ell < t_i$, where T_ℓ denotes the T_{MRCA} at locus ℓ (and likewise for $\mathcal{L}_\ell$). The set of possible states $\{s_1, ..., s_S\}$ is denoted by $\mathcal{S}$. The sequence of states the CHMM occupies along the complete genome is $\vec{s} = (s^1, \ldots, s^L)$, where s^ℓ is the state at locus ℓ.

For the emission observed at a given locus, our method uses the number of derived alleles d at that locus. Since the data consists of n haplotype sequences, we can observe up to $n-1$ derived alleles at a locus, thus the set of possible emissions is $\mathcal{D} := \{0, ..., n-1\}$. Note that $\mathcal{D}$ includes 0 to model loci where all samples share the same allele. The vector of observations across the genome is $\vec{d} = (d^1, \ldots, d^L\}$, where d^ℓ is the number of derived alleles at locus ℓ. With these definitions for the state space and emission space for our CHMM, we introduce the transition and emission probabilities, given by matrices $\mathbf{A}$ and $\mathbf{B}$, respectively, with elements

$$A_{ij} = \mathbb{P}[s^{\ell+1} = s_j | s^\ell = s_i], \tag{2.1}$$

$$B_{id} = \mathbb{P}[d^\ell = d | s^\ell = s_i], \text{ and} \tag{2.2}$$

$$\Pi_i = \mathbb{P}[s^\ell = s_i]. \tag{2.3}$$

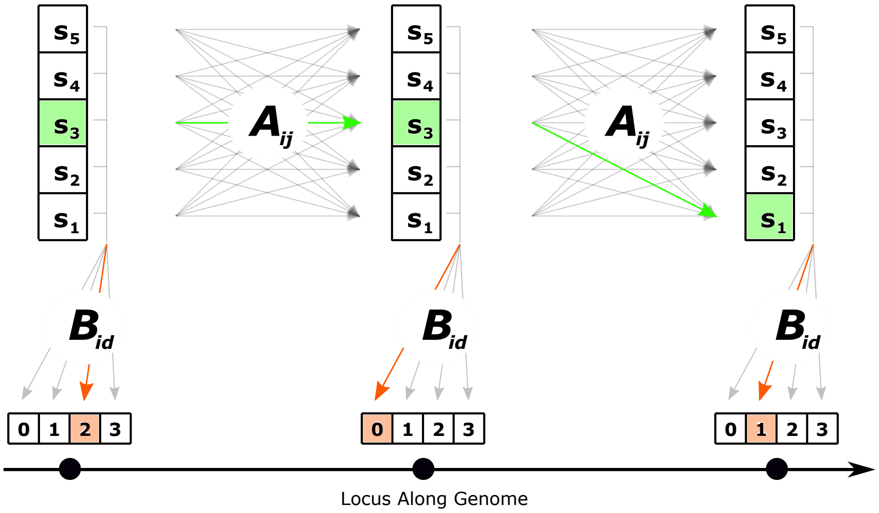


Figure 2.2: Schematic of our CHMM for a sample of size 4. Information about the underlying tree at each locus is captured by the state s_i. $\mathcal{S} = \{s_1, \ldots, s_5\}$ is the set of intervals into which the respective summary of the tree (T_{MRCA} or $\mathcal{L}$) can fall. The states change from each locus to the next in accordance with the transition matrix $\mathbf{A}$ and the observed number of derived alleles at each locus is emitted in accordance with the emission probabilities $\mathbf{B}$.

The quantity $\mathbf{\Pi}$ is the marginal distribution of the hidden states, and thus it is also the distribution of s^0, the first state in the CHMM. Figure 2.2 depicts a schematic of the transition and emissions in this CHMM.

To summarize, the essential elements of our model are the sequence of latent states along the genome $\vec{s}$, the observed emission sequence $\vec{d}$, the transition matrix $\mathbf{A}$, the emission matrix $\mathbf{B}$, and the marginal distribution of hidden states $\mathbf{\Pi}$. In the next few sections, we show how we implement this model. In section 2.3 and section 2.4 we show how to compute entries of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{\Pi}$ based on the population history $\eta(t)$ and the biological parameters ρ and θ. Once these probabilities are computed, we can infer the hidden state sequence given the observed data and compute the posterior likelihood.

2.3 Computing Probabilities with T_{MRCA}

To use T_{MRCA} as the hidden state in our CHMM, we discretize the continuous random variable into discrete intervals partitioned at certain t_i as previously described in Section 2.2.

We now describe the analytic and numerical methods used to compute the corresponding transition and emission probabilities in equation (2.1), (2.2), and (2.3). This is a critical step in our model because it links the parameters of demographic history of the population to the key quantities of our CHMM. Briefly, we first define an augmented version of the ancestral process with recombination (for transitions) or with mutation (for emissions). These ancestral processes are modifications of the regular coalescent process that allow the modeling of recombination events between loci and mutation events. The distribution of these continuous-time Markov processes can be obtained by solving the associated system of ODEs defined by the respective rate matrices. We numerically obtain the solutions for these ODEs and combine them appropriately to compute the transition and emission probabilities of the CHMM. The mutation and recombination rates are given in units of coalescent-time as before, denoted by θ and ρ, respectively.

2.3.1 Transition Probabilities

To compute the transition probabilities, we employ an augmented ancestral process with recombination, $\mathcal{A}^{\rho}$, introduced by [21]. This process closely resembles the regular ancestral process with recombination described by [26] and describes the joint distribution of the genealogies of n samples for two adjacent loci separated by a recombination distance of ρ. We use $\mathcal{A}^{\rho}$ to compute the respective transition probabilities in the matrix $\mathbf{A}$.

The process $\mathcal{A}^{\rho}(t)$ is initialized at the present ($t = 0$) and tracks the ancestral lineages at two loci, a and b, simultaneously as they evolve backwards in time. Initially, there are n lineages, each ancestral to both loci a and b of one of the n sampled haplotypes. As in the standard coalescent with varying population size, ancestral lineages coalesce with rate $\lambda(t)$. Additionally, recombination events can occur on each lineage ancestral to two loci at rate $\frac{\rho}{2}$ and decouples the dynamics of the two loci by splitting the respective lineage into two lineages that are ancestral at only a single locus. The two decoupled lineages then evolve

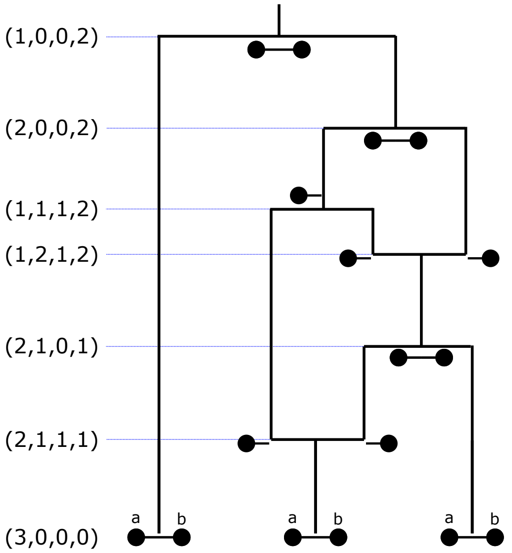


Figure 2.3: Example trajectory of $\mathcal{A}^\rho(t)$ with the state denoted by tuples. At $t = 0$ there are three lineages of type k_{ab} ancestral to both loci for their respective samples. The lineages split at ancestral recombination events and combine at coalescence events where they find a common ancestor. The trajectory ultimately culminates in the state $(1, 0, 0, 2)$, signifying that there is one lineage ancestral to both loci in all present-day samples, and that two recombination events occurred in this genealogy.

independently and yield distinct genealogies for each of the loci. Ultimately all lineages coalesce into a common ancestor for both loci.

The states of $\mathcal{A}^\rho(t)$ are denoted by (k_{ab}, k_a, k_b, r) which are tuples describing the configuration of lineages. Here, k_{ab} is the number of active lineages ancestral to both loci (coupled), k_a are the lineages ancestral only to locus a, and k_b are the lineages ancestral only to b (uncoupled). Finally, r explicitly tracks the number of recombination events that have occurred since $t = 0$. While $k_{ab} \in \{1, 2, ..., n\}$, the decoupled lineages are constrained such that $k_a, k_b \in \{0, 1, ..., r\}$ since there can be at most as many uncoupled lineages as recombination events. In the full ancestral process, r takes values from 0 to ∞, since there can be an arbitrary number of recombination events between the two loci.

Figure 2.3 shows an example trajectory of this augmented ancestral process. Recombination events decouple a shared lineage, while coalescence events fuse two active lineages. If an uncoupled lineage (one of type k_a or type k_b) coalesces with a coupled lineage (one of type

k_{ab}), the resulting lineage contains ancestry that traces to both loci a and b in our sample, and is therefore a coupled lineage. We note that $\mathcal{A}^\rho$ allows for a joint lineage to recombine (split) and immediately coalesce back together making this process consistent with the SMC' formulation [27], which is a more faithful model of the true genetic process than the original SMC formulation [11].

The unboundedness of r renders this full process difficult to solve. In the remainder of this work, we restrict r to be at most 1 (and consequently also restrict k_a and k_b to be 0 or 1). This restriction is an approximation that is accurate when the recombination rate between two adjacent loci is low. In such a case, we expect to see at most one ancestral recombination event separating the genealogies at two neighboring loci. This approximation is justified because in many real world organisms, such as humans, the recombination rate is low. Henceforth, $\mathcal{A}^\rho(t)$ will refer to this restricted process.

We set $\mathcal{A}^\rho(0) = (n, 0, 0, 0)$, which corresponds to initializing the process in a state with n lineages at the present time, each ancestral to both loci in a single sample. The absorbing states are $(1, 0, 0, 1)$ and $(1, 0, 0, 0)$, which correspond to the states where all lineages for both loci have fully coalesced after 0 or 1 recombination events have occurred.

The possible transitions between states and their respective rates are given in Table 2.1. The rates in the first three rows of this table correspond to all possible coalescent events. These rates are all proportional to the time-dependent coalescent rate $\lambda(t)$. The first row describes coalescence among the lineages ancestral to locus a and b, which results in reducing the number of these lineages by 1. The rate for these events is proportional to all possible pairs of such lineages. The second row describes events that reduce the number of lineages ancestral to only a by one, which can either be a coalescent event among the a lineages, or a coalescence event between one lineage ancestral to a and another ancestral to a and b. Again, the rate is proportional to the number of such lineage pairings. The third row describes the respective events for the b lineages. The fourth row corresponds to recombination events,

Transition from (k_{ab}, k_a, k_b, r) to:	Rate
$(k_{ab}-1, k_a, k_b, r)$	$\lambda(t)\binom{k_{ab}}{2}$
(k_{ab}, k_a-1, k_b, r)	$\lambda(t)\left[\binom{k_a}{2}+k_a k_{ab}\right]$
(k_{ab}, k_a, k_b-1, r)	$\lambda(t)\left[\binom{k_b}{2}+k_b k_{ab}\right]$
$(k_{ab}-1, k_a+1, k_b+1, r+1)$	$\begin{cases} k_{ab}\frac{\rho}{2}, & \text{if } r=0 \\ 0, & \text{else} \end{cases}$

Table 2.1: Possible transitions from a given state (k_{ab}, k_a, k_b, r) and their respective rates. The first row gives the rate for coalescence between two lineages that are ancestral to both loci. The second row gives rate for two types of events, coalescences between two lineages ancestral to only locus a, and coalescences of a lineage ancestral only to a with a lineage ancestral to both. The third row reflects similar events for locus b. The last row gives the rate of recombination events. Note that these rates are defined to permit a maximum of 1 ancestral recombination event occurring between locus a and b.

with a rate proportional to the recombination rate $\rho/2$. These recombination events can only happen in lineages ancestral to a and b and reduces their number by one. Such events result in one lineage ancestral to only a and one only to b, increasing the respective numbers by one, and also increasing r by one. Note that the rates for the recombination events reflect the fact we restrict the process to have at most one recombination event.

Define $g_\sigma^\rho(t) := \mathbb{P}[\mathcal{A}^\rho(t) = \sigma]$ to be the probability that the augmented ancestral process is in state $\sigma \in \mathcal{R}$ at time t, where $\mathcal{R}$ is the set of all possible states. Then $\vec{g}^\rho(t) = \left(g_\sigma^\rho(t)\right)_{\sigma\in\mathcal{R}}$ is the distribution of the process at time t, a vector of probabilities over all the states $\sigma \in \mathcal{R}$. Since $\mathcal{A}^\rho$ is a continuous-time Markov process, the evolution of $\vec{g}(t)$ is given by the system of ordinary differential equations (ODEs)

$$\frac{d}{dt}\vec{g}^\rho(t) = \vec{g}^\rho(t)\cdot\mathbf{Q}^\rho(t), \tag{2.4}$$

where $\mathbf{Q}^\rho(t)$ is the rate matrix consisting of the rates given in Table 2.1. The rate matrix is time-dependent, since the coalescent rates $\lambda(t)$ are as well. We can now obtain the probabilities $\vec{g}(t)$ by numerically integrating equation (2.4) (see 2.6.2 for details).

Moreover, from the distribution $\vec{g}(t)$, we can compute the cumulative joint distribution function (CDF) of the T_{MRCA} at the two loci, $\mathbb{P}[T_a \leq \tau_a; T_b \leq \tau_b]$, where $0 \leq \tau_a, \tau_b < \infty$, and T_a and T_b are the T_{MRCA}'s at a and b respectively. Without loss of generality, we assume that $\tau_a < \tau_b$ and obtain

$$\begin{aligned}
\mathbb{P}[T_a \leq &\tau_a; T_b \leq \tau_b] \\
&= \mathbb{P}[T_a \leq \tau_a; T_b \leq \tau_a] + \mathbb{P}\big[T_a \leq \tau_a; T_b \in (\tau_a, \tau_b]\big] \\
&= g_{(1,0,0,0)}(\tau_a) + g_{(1,0,0,1)}(\tau_a) + \mathbb{P}\big[T_a \leq \tau_a; T_b \in (\tau_a, \tau_b]\big] \\
&= g_{(1,0,0,0)}(\tau_a) + g_{(1,0,0,1)}(\tau_a) + g_{(1,0,1,1)}(\tau_a) \cdot \\
&\qquad\qquad \mathbb{P}\big[\sigma(\tau_b) = (1,0,0,1) \mid \sigma(\tau_a) = (1,0,1,1)\big] \\
&= g_{(1,0,0,0)}(\tau_a) + g_{(1,0,0,1)}(\tau_a) + g_{(1,0,1,1)}(\tau_a) \cdot \big[1 - e^{-\int_{\tau_a}^{\tau_b} \lambda(t) dt}\big]. \qquad (2.5)
\end{aligned}$$

In the first equality, we partition the probability according to whether $T_b < \tau_a$ or $T_b > \tau_a$. The second equality holds, because $\mathbb{P}[T_a \leq \tau_a; T_b \leq \tau_a]$ is the probability that the ancestral process is in an absorbing state where both loci have found the T_{MRCA} by time τ_a. The third equality follows from the fact that $\mathbb{P}\big[T_a \leq \tau_a; T_b \in (\tau_a, \tau_b]\big]$ is the probability that the lineages at a found a common ancestor by time τ_a and the lineages at b find a common ancestor after τ_a, but before τ_b, which is only possible if a recombination event occurred. Since this term is conditional on a having found its T_{MRCA}, only 2 lineages can be remaining (one ancestral only to b, and one the common ancestor of a) due to the assumption that $r \leq 1$, and thus the term simplifies to the coalescence probability of two lineages between times τ_a and τ_b. The final equality follows.

By evaluating equation (2.5) at the values $\tau_a, \tau_b \in \{t_i\}_{i=0}^{S}$, the interval boundaries for the discretized state space, we can obtain a joint cumulative distribution matrix for T_a and T_b, denoted $\mathbf{A}^{CDF}$. From this it is straightforward to compute the matrix of joint probabilities, $\mathbf{A}^{PDF}$. Dividing the joint probabilities by the marginal probabilities, we arrive at $\mathbf{A}$, the transition probability matrix itself:

$$A_{ij}^{CDF} := \mathbb{P}[T_a \leq t_i; T_b \leq t_j]$$

$$A_{ij}^{PDF} := A_{ij}^{CDF} - A_{i-1,j}^{CDF} - A_{i,j-1}^{CDF} + A_{i-1,j-1}^{CDF} \tag{2.6}$$

$$A_{ij} := \frac{A_{ij}^{PDF}}{\sum_{k \in \mathcal{S}} A_{ik}^{PDF}} \tag{2.7}$$

Additionally, we can obtain the vector of marginal probabilities $\mathbf{\Pi}$ as

$$\Pi_i = \sum_{j \in \mathcal{S}} A_{ij}^{PDF}. \tag{2.8}$$

2.3.2 Emission Probabilities

To compute the emission probabilities with T_{MRCA} as the hidden state, we introduce an augmented single-locus ancestral process with mutation $\mathcal{A}^{\theta}$ which is an extension to the regular ancestral process [23] that is motivated by the fact that, conditional on the coalescent tree, mutations are Poisson distributed along the branches with rate $\frac{\theta}{2}$. Similar to the recombination case, we only consider at most one mutation event (motivated by the assumption that the per locus mutation rate is low). The states of this process are denoted by (k, k^*), where k is the number of active lineages ancestral to the n samples, and k^* is the number of lineages that were active at the time of the first mutation event along the genealogy (going backwards in time). If no mutation has occurred yet, k^* assumes a value of -1.

The process is initialized in $(n, -1)$, a state before any mutation has occurred with one ancestral lineage for each sample. The transition rates are given in Table 2.2 and an example trajectory is shown in Figure 2.4. The possible transitions are either two lineages coalescing

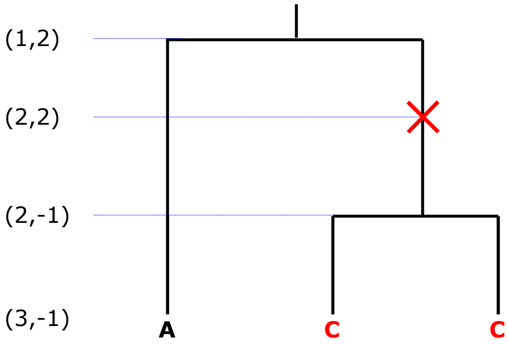


Figure 2.4: Example trajectory of the ancestral process with mutation for $n = 3$ samples with the state (k, k^*) indicated on the left. The mutation process is superimposed onto the regular genealogical process. In this example, the mutation happens when there are two ancestral lineages, resulting in two samples carrying the derived allele.

or a lineage mutating. The rate for coalescence is given by the coalescent rate $\lambda(t)$ times the number of possible pairs that can coalesce, and such an event reduces the number of lineages by one. The rate for a transition via mutation is given by the mutation rate $\frac{\theta}{2}$ multiplied with the number of lineages that a mutation can occur on. The number of lineages that were active at the time of the mutation event is recorded in the second component of the state. Since we restrict to one mutation event at most, if this number is set once, it will not be set again. The process is absorbed in any state $(1, k^*)$ with $k^* \in \{-1, 2, \dots, n\}$. Note that it is important to continue the process after a mutation event occurred until all lineages are coalesced so that we obtain the full distribution of the T_{MRCA}.

Transition	Rate
$(k, k^*) \to (k-1, k^*)$	$\lambda(t)\binom{k}{2}$
$(k, -1) \to (k, k)$	$k\frac{\theta}{2}$

Table 2.2: The transition rates of the augmented ancestral process $\mathcal{A}^\theta$. The first row gives the rate of a coalescence event of two lineages, while the second row gives the rate for mutation events. Note that only one mutation event is permitted.

Similar to the procedure for $\mathcal{A}^\rho$, we collect all transition rates in a matrix $\mathbf{Q}^\theta(t)$. We further define $g^\theta_\sigma(t) := \mathbb{P}[\mathcal{A}^\theta(t) = \sigma]$ as the probability that the ancestral process $\mathcal{A}^\theta$ is in a state $\sigma \in \mathcal{M}$ at time t, where $\mathcal{M}$ is the set of all possible states. The evolution of the vector of all probabilities $\vec{g}^{\theta}(t) = \left(g^{\theta}_{\sigma}(t)\right)_{\sigma \in \mathcal{M}}$ is given by the system of ODEs

$$\frac{d}{dt}\vec{g}^{\,\theta}(t) = \vec{g}^{\,\theta}(t) \cdot \mathbf{Q}^{\theta}(t).$$

Again, we obtain the solution to these ODEs numerically as described in 2.6.2.

Using the distribution of this process, we can compute the cumulative distribution of T_{MRCA} jointly with the probability of emitting d derived alleles as

$$\mathbb{P}[T_{MRCA} \leq \tau; 0 \text{ derived alleles}] = \mathbb{P}[\mathcal{A}^{\theta}(\tau) = (1, -1)],$$

since this gives the probability that all lineages are coalesced by τ and no mutation occurred, and

$$\begin{aligned}
&\mathbb{P}[T_{MRCA} \leq \tau; d \text{ derived alleles}] \\
&\qquad = \mathbb{P}[\mathcal{A}^{\theta}(\tau) = (1, k^*) \text{ for } k^* \in \{2, \dots, n\}; d \text{ derived alleles}] \\
&\qquad = \sum_{k^* \in \{2,\dots,n\}} g^{\theta}_{(1,k^*)}(\tau) \cdot \mathbb{P}\big[d \text{ derived alleles} | \text{mutation while } k^* \text{ lineages}\big] \\
&\qquad = \sum_{k^* \in \{2,\dots,n\}} g^{\theta}_{(1,k^*)}(\tau) \cdot \frac{\binom{n-d-1}{k^*-2}}{\binom{n-1}{k^*-1}} \qquad (2.9)
\end{aligned}$$

for $d \in \{1, \dots, n-1\}$. The first equality in equation (2.9) follows from the fact that $T_{MRCA} < \tau$ if and only if the ancestral process has found an absorbing state (all lineages have coalesced) before τ. In the second equality, we partition this probability with respect to the specific number of lineages active when the mutation occurred, which is encoded in the absorbing state. For the last equality, we substitute the probability of emitting a certain number of derived alleles given that there were k^* active lineages at the time of the mutation. This probability is given by the probability that one of the k^* lineages subtends d leafs [e.g. 28, Ch. 2.1]. It is independent of the time of the mutation.

By evaluating these probabilities at times $\tau \in \{t_i\}_{i=0}^S$, we compute the discretized joint CDF for the emissions, $\mathbf{B}^{CDF}$, which is again used to compute the joint probabilities $\mathbf{B}^{PDF}$ and ultimately the emission probabilities $\mathbf{B}$ for the CHMM by conditioning on the hidden state:

$$B_{id}^{CDF} := \mathbb{P}[T_{MRCA} \leq t_i; y = d] \tag{2.10}$$

$$B_{id}^{PDF} := B_{id}^{CDF} - B_{i-1,d}^{CDF}$$

$$B_{id} := \frac{B_{id}^{PDF}}{\sum_k B_{ik}^{PDF}}. \tag{2.11}$$

2.4 Computing Probabilities with $\mathcal{L}$

We now describe the implementation of our CHMM with the total branch length (sum of all branch lengths) of the genealogical tree $\mathcal{L}$ as the hidden state at each locus. As before, we discretize $\mathcal{L}$ by partitioning the real line with a set of values $t_0 = 0, < t_1 < \ldots < t_S = \infty$, and say that the CHMM is in state s_i at the given locus, if $t_{i-1} \leq \mathcal{L} < t_i$. In order to compute the joint distributions for $\mathbf{A}^{CDF}$ and $\mathbf{B}^{CDF}$, we employ the approach first outlined in [21] using partial differential equations (PDEs) associated with the augmented ancestral processes $\mathcal{A}^\rho$ and $\mathcal{A}^\theta$. We numerically solve the respective systems and use the solutions to obtain the transition and emission probabilities as follows.

2.4.1 Transition Probabilities

We follow the approach in [21] to compute the joint distribution of the marginal total tree length at locus a and b. We begin by computing the joint distribution of the total tree length accumulated up to a certain time t in the past. Using both the augmented ancestral process

$\mathcal{A}^\rho$ (introduced in Section 2.3.1) to compute the requisite distributions for T_{MRCA} and

$$v^\ell(k_{ab}, k_a, k_b, r) := (k_{ab} + k_\ell)\mathbb{1}_{\{k_{ab}+k_\ell>1\}}, \tag{2.12}$$

we can define

$$A_a(t) = v^a\big(\mathcal{A}^\rho(t)\big)$$

and

$$A_b(t) = v^b\big(\mathcal{A}^\rho(t)\big)$$

to count the number of active lineages that are ancestral to locus a or b at a given time t. Note that this includes the lineages ancestral to both loci. We define the marginally accumulated tree length

$$L_\ell(t) := \int_0^t A_\ell(s)ds. \tag{2.13}$$

The quantities $L_a(t)$ and $L_b(t)$ can be thought of as the total branch lengths that has aggregated at each locus as the process evolves back in time. This holds because the integrand in equation (2.13) is the number of lineages at a specific locus at a given time and total branch length is accumulated linearly along each active lineage. The indicator function in equation (2.12) signifies that the process stops accumulating tree length once only a single lineage is left, ie. the T_{MRCA} is reached. Using this notation, we can define the probabilities

$$F_\sigma(t, x, y) := \mathbb{P}\big[\mathcal{A}^\rho(t) = \sigma, L_a(t) \le x, L_b(t) \le y\big]$$

which give the joint distribution of tree length accumulated at both loci up to time t and of the ancestral process $\mathcal{A}^\rho(t)$ being in state σ

The values of $\vec{F} := \big(F_\sigma(t, x, y)\big)_{\sigma\in\mathcal{R}}$ can be obtained as solutions to the system of PDEs

$$\partial_t\vec{F} + \partial_x\vec{F} \cdot \mathbf{V}^a + \partial_y\vec{F} \cdot \mathbf{V}^b = \vec{F} \cdot \mathbf{Q}^\rho(t) \tag{2.14}$$

and its corresponding boundary conditions

$$F_\sigma(t,x,y) = \begin{cases} \mathbb{P}\big[\mathcal{A}^\rho(t) = \sigma, L_b(t) \leq y\big], & \text{if } x \leq nt \\ \mathbb{P}\big[\mathcal{A}^\rho(t) = \sigma, L_a(t) \leq x\big], & \text{if } y \leq nt \\ 0, & \text{if } x = 0 \text{ or } y = 0 \end{cases} \tag{2.15}$$

as shown in [21]. These equations are given in terms of the rate matrix $\mathbf{Q}^\rho(t)$ of the augmented ancestral process $\mathcal{A}^\rho$ and the diagonal matrices

$$\mathbf{V}^\ell := \operatorname{diag}\{v^\ell(\sigma)\},$$

which represent the accumulation of tree length along the active ancestral lineages.

In [21], the authors show that the quantities $\mathbb{P}\big[\mathcal{A}^\rho(t) = \sigma, L_a(t) \leq x\big] =: F_\sigma(t,x)$ (and the corresponding quantities for b) can in turn be obtained as the solution of the PDEs

$$\partial_t \vec{F} + \partial_x \vec{F} \cdot \mathbf{V}^a = \vec{F} \cdot \mathbf{Q}^\rho(t) \tag{2.16}$$

with boundary conditions

$$F_\sigma(t,x) = \begin{cases} \mathbb{P}\big[\mathcal{A}^\rho(t) = \sigma\big] = g^\rho_\sigma(t), & \text{if } x \leq nt \\ 0, & \text{if } x = 0. \end{cases}$$

We implemented the scheme outlined in [21, Appendix B] to compute the solutions to these PDEs and provide additional details of our implementation in Section 2.6.3.

Lastly, the joint distributions of tree length at loci a and b can be obtained from the solutions of the absorbing states of $\mathcal{A}^\rho$ and are given by

$$\mathbb{P}[\mathcal{L}_a \leq x, \mathcal{L}_b \leq y] = \Big[F_{(1,0,0,0)}(t,x,y) + F_{(1,0,0,1)}(t,x,y)\Big]\Big|_{t=\frac{max(x,y)}{2}}, \tag{2.17}$$

where $\mathcal{L}_a$ and $\mathcal{L}_b$ denote the total branch length of the genealogies at locus a and b respectively [21]. Evaluating these probabilities at $x, y \in \mathcal{S} = \{t_0, t_1, \ldots, t_S\}$ yields the elements of the joint cumulative probability matrix $\mathbf{A}^{CDF}$. This discretized joint distribution can then be used in equations (2.6) and (2.7) to compute $\mathbf{A}^{PDF}$ and ultimately the transition probabilities $\mathbf{A}$ for the CHMM when using the total tree length $\mathcal{L}$ as the hidden state. Similarly, the initial distribution can be obtained using equation (2.8).

2.4.2 Emission Probabilities

Computing the emission probabilities closely follows the steps for the transition probabilities outlined in the previous section. However, we use the ancestral process with mutation $\mathcal{A}^\theta$ instead of the process with recombination $\mathcal{A}^\rho$, and instead of one variable for time and two for tree length (t, x, y), we only need to use one variable for time and one for tree length (t, x) since we only consider emission at one locus. Before we can compute the emission probabilities, we first need to compute the joint probability of accumulating a certain tree length by t and $\mathcal{A}^\theta(t)$ occupying a certain state:

$$F_{(k,k^*)}(t,x) = \mathbb{P}\big[\mathcal{A}^\theta(t) = (k,k^*), L(t) \leq x\big].$$

Here, $L(t)$ is the accumulated tree length at this locus, defined similarly to equation (2.13) as

$$L(t) := \int_0^t v^\theta\big(\mathcal{A}^\theta(s)\big)ds,$$

where $v^{\theta}(k,k^*) = k\mathbb{1}_{\{k>1\}}$. Similar to Section 2.4.1 and [21], the vector of these probabilities can be obtained as the solution to the following system of PDEs

$$\partial_t \vec{F} + \partial_x \vec{F} \cdot \mathbf{V}^{\theta} = \vec{F} \cdot \vec{Q}^{\theta}(t), \tag{2.18}$$

with boundary conditions

$$F_{(k,k^*)}(t,x) = \begin{cases} \mathbb{P}\big[\mathcal{A}^{\theta}(t) = (k,k^*)\big] g^{\theta}_{(k,k^*)}(t), & \text{if } x \leq nt \\ 0, & \text{if } x = 0, \end{cases}$$

where $\vec{Q}^{\theta}(t)$ is the matrix of transition rates of the process $\mathcal{A}^{\theta}$ and the diagonal matrix

$$\mathbf{V}^{\theta} := \operatorname{diag}\{v^{\theta}(\sigma)\}.$$

Solving this system is similar to solving the PDEs in Section 2.4.1, and our method is described in Section 2.6.3. Similar to equation (2.9), we can then combine the probabilities for the absorbing states with the respective combinatorial factors to obtain the joint probability distribution of the tree length $\mathcal{L}$ and the observed number of derived alleles as

$$\mathbb{P}[\mathcal{L} \leq x; y = d] = \sum_{k^*} F^{\theta}_{(1,k^*)}(t,x)\Big|_{t=\frac{x}{2}} \cdot \frac{\binom{n-d-1}{k^*-2}}{\binom{n-1}{k^*-1}}.$$

We can then again evaluate these probabilities at $x \in \mathcal{S} = \{t_0, t_1, \ldots, t_S\}$, the discretization points of the $\mathcal{L}$ states, to obtain the entries of the matrix of cumulative probabilities $\mathbf{B}^{CDF}$, which can be substituted into equation (2.10) and (2.11) to obtain $\mathbf{B}^{PDF}$, and ultimately the emission probabilities $\mathbf{B}$ for the CHMM using $\mathcal{L}$ as the hidden state, that is, the probabilities of observing a certain number of derived alleles, given the tree length.

2.5 Computing Likelihoods and Inferring Hidden States

In this section, we describe how we infer the hidden state sequence $\vec{s}$ given the observed emission sequence $\vec{d}$, and how to compute the likelihood of observing $\vec{d}$ given the model parameters ($\mathbf{A}$, $\mathbf{B}$, and $\mathbf{\Pi}$). We also describe augmentations we make to the standard HMM machinery in order to improve the efficiency. These include strategies to improve the speed of the forward-backwards algorithm, and a composite-likelihood model that allows efficient analysis of large sample sizes. This section contains some details about our specific implementation of CHIMP that go beyond the theory and models.

2.5.1 The Forward-Backward Algorithm

The sequence of hidden states for our CHMM can be inferred by applying the Forward-Backward algorithm [29, Ch. 13.2.2] to the observed genotype data $\vec{d} = (d_1, \ldots, d_L)$, where d_ℓ is the number of derived alleles at locus ℓ. This is a well known algorithm which first traverses the emission sequence forwards and then backwards, computing intermediate probabilities that are ultimately combined to compute the posterior distribution across states for each locus,

$$\gamma_i(\ell) = \mathbb{P}[s^\ell = s_i | \vec{d}], \tag{2.19}$$

and $\mathbb{P}[\vec{d}]$, the likelihood of observing data given the specified model. The scaled version of the Forward-Backward algorithm [29, Ch. 13.2.4] allows for greater numerical stability since the small probabilities involved in the standard algorithm can be much smaller than machine precision.

Additionally, performing computations at each nucleotide site in the genome can be computationally prohibitive, so we can employ one of two strategies to speed the Forward-Backward computations up: a *locus-skipping* method detailed in 2.5.2 and a *meta-locus* model detailed in 2.5.3.

2.5.2 Locus-Skipping

The *locus-skipping* method to improve efficiency of the Forward-Backward algorithm was first described in the supplement of [16], and our implementation follows this blueprint. This is a modification to the forward and backward algorithm that computes the exact likelihood for tracts of monomorphic sites in a single-step, thereby greatly improving efficiency. Briefly, this method uses the eigendecomposition of the single-step transition matrix at monomorphic sites to efficiently exponentiate the respective matrix and integrate large tracts of monomorphic sites in a single step. It is important to note that the underlying likelihood model is exact and equal to the model that considers each site individually. Thus, this method improves the efficiency of the Forward-Backward algorithm with no trade-offs in accuracy. We also note that for the locus-skipping step some of the probabilities are stored as *log* values to avoid machine precision issues. Even with this measure numerical errors arise when the tract being skipped is too large, and in practice we cap the size of these skips at 1000 loci. The two types of monomorphic tracts that our implementation can handle are tracts of non-segregating sites where there is no genetic variation across the samples, and tracts of missing data where the sequencing process is inconclusive. These types of tracts tend to appear commonly in genomic data, so using the locus-skipping method shows clear payoff.

2.5.3 Meta-Locus Model

The *meta-locus* model is an alternative to the *locus-skipping*, and involves a slight modification of the underlying likelihood model itself, following a strategy that has been described in [18, Suppl. Text 4.2]. Rather than implementing the CHMM with a hidden state and an emission at each nucleotide site, we combine a specified number of sites into a meta-locus,

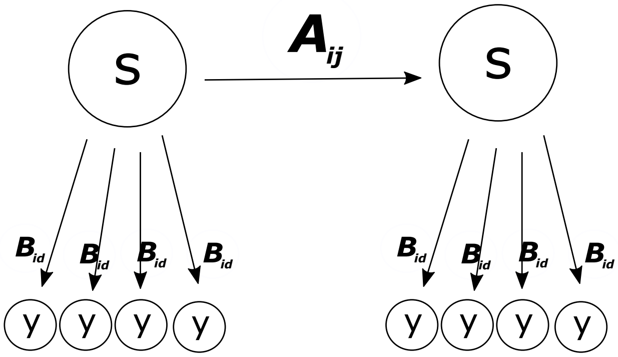


Figure 2.5: Schematic for *meta-locus* model with $s = 4$. Each meta-locus represents 4 sites. Recombination does not happen within a meta-locus, but between at elevated rates. The emission probabilities $\mathbf{B}_{id}$, 4 for each meta-locus, are the same as in the single-step model.

with a single hidden state but with several emissions for all nucleotides that are grouped into this meta-locus. We specify in advance the number of nucleotide sites s that a meta-locus spans, which also determines the number of emissions per locus. Figure 2.5 shows a schematic of this approximation.

The emission probabilities for each site in this model are identical to those used in the full model. However, by assuming a single hidden state for all sites in a meta-locus, we effectively suppress recombination between these, and thus it is necessary to implement recombination between two meta-loci at correspondingly elevated rates. To this end, the transition probabilities are obtained by raising the single-site transition matrix to the power s (the number of sites in the meta-locus).

A potential drawback of this strategy is that linkage information is lost within each metal-locus. However, using meta-loci also offers the potential for a tremendous speedup of the Forward-Backward algorithm and is extremely useful to offset additional computational costs that are incurred when using large numbers of CHMM states. We find that, in general, there is a sweet spot where the gains in efficiency are significant and the loss of information is negligible.

2.5.4 Composite Likelihood

Another feature of our likelihood model is the way in which we handle large sample sizes. As input, our method takes genotype data of a sample of n haploids at L consecutive sites of the genome. The numerical procedures to compute the transition and emission probabilities are computationally more expensive for larger sample sizes, so to efficiently scale up, we use a composite likelihood scheme that partitions the full collection of n samples into a number of non-overlapping sub-groups, each of size n_s. The genomic data is then split into separate sequences, $\vec{d}^{(i)}$ where the superscript i here indexes the subgroups. The Forward-Backward algorithm is then performed for each $\vec{d}^{(i)}$ independently and the contributions of each subgroup are added together to obtain the composite likelihood

$$\sum_{i=1}^{\lfloor \frac{n}{n_s} \rfloor} \log \mathbb{P}[\vec{d^{(i)}}].$$

Given our implementation of the composite likelihood model, the complexity of the Forward-Backward algorithm scales linearly with number of samples (doubling the sample size doubles the number of independent genomic tracts spanning n_s samples), although in principle each subgroup could be processed in parallel to make the run-time invariant to sample size.

Since the composite likelihood model equips our method to handle multiple independent contiguous genomic sequences spanning n_s samples each, data from multiple chromosomes can straightforwardly be handled in the same analysis by treating them as independent sequences.

2.6 Additional Implementation Details

2.6.1 Choosing Interval Boundaries for Hidden States

For both T_{MRCA} and $\mathcal{L}$, we have the task of deciding how to partition the continuous values for these quantities into discrete states to use for the CHMM. A possible strategy would be to divide the continuum into equidistant intervals on a logarithmic or geometric scale [e.g. 13]. However, here we choose a different strategy and divide the continuum into states such that the marginal probabilities of the CHMM occupying any one state at a given locus are approximately uniform. We compute this partition for the uniform distribution in a model of constant population size, where we estimate the size from the number of segregating sites in the given sample using Watterson's estimator (fixing the per generation mutation rate). This causes each state to have comparable weight in the analysis, and avoids having infrequently-visited states that contribute little information. CHMM methods gather much of their power from observing the various transitions between the different states, since these are informative in estimating off-diagonal elements of the transition matrix which contain much of the signal used to distinguish between population size histories during estimation.

In order to partition the continuum, we numerically partition it so that the distribution of T_{MRCA} or $\mathcal{L}$ is uniform across the states. The analytical formulas for the distributions of these quantities in the case of a constant population size are

$$f_{T_{\mathrm{MRCA}}}(t) = \sum_{i=2}^{n} \frac{n(n-1)\cdots(n-i+1)}{n(n+1)\cdots(n+i-1)}(2i-1)(-1)^i \binom{i}{2} e^{-\binom{i}{2}t}$$

and

$$f_{\mathcal{L}}(t) = \frac{n-1}{2} e^{-t/2}\left(1-e^{-t/2}\right)^{n-2},$$

see for example equations (3.28) and (3.34) of [30]. By numerically evaluating these distributions, we find the respective uniform partitions.

2.6.2 Numerically Integrating ODEs

While computing transition and emission probabilities, for both T_{MRCA} and $\mathcal{L}$, we numerically have to solve systems of ODEs. For T_{MRCA}, the solutions of the ODE are directly used in computing the probabilities (equation (2.5) and equation (2.9)), whereas for $\mathcal{L}$ the solutions to the ODEs are used in defining the boundary conditions of the PDE problem that has to be solved (equation (2.15)). For all these applications, we use the Apache Commons' implementation of a Dormand-Prince algorithm of order 8(5,3) [31].

2.6.3 Solving PDEs for $\mathcal{L}$

To numerically compute the requisite solutions to the PDEs (2.14) and (2.18), we implement the solution scheme outlined in [21]. We will give a brief overview of the scheme here, but refer the reader to the original publication for more details. Moreover, we will focus on the solution scheme for the joint distribution of the tree length $F_\sigma(t, x, y)$ in equation (2.14) here, but a similar scheme can be used for the emission distribution (2.18).

In this scheme, $F_\sigma(t, x, y)$ is computed for each state sequentially, beginning with the initial state $(n, 0, 0, 0)$. Because we explicitly keep track of the number of recombination events, we can ensure that each transition of the associated Markov chain either increases the number of recombination events or decreases some number of lineages. The states are thus ordered, and following this order in the evaluation, we can ensure that when computing $F_\sigma(t, x, y)$, all preceding $F_{\sigma'}(t, x, y)$ that are necessary for the respective computations have already been computed. In other words, since there are no loops in the probability flow through states, we can solve them sequentially following the natural ordering.

To compute the solution for a specific $F_\sigma(t, x, y)$, we favor a bespoke implementation of the PDE solutions as opposed to using a black box numerical PDE solver, since the actual solutions for each state are computed using the Method of Characteristics. This allows us to take advantage of partially analytic results which can reduce numerical errors and increase efficiency. Moreover, in terms of efficiency, we believe solving the states sequentially is faster and more accurate than attempting to simultaneously solve for all states using a general purpose method.

For each state $\sigma = (k_{ab}, k_a, k_b, r)$ we then define a three dimensional grid of points for the function $F_\sigma(t, x, y)$, with $0 \leq t \leq \frac{t_{S-1}}{2}$, $v^a(\sigma) \cdot t \leq x \leq n \cdot t$, and $v^b(\sigma) \cdot t \leq x \leq n \cdot t$. The grid consists of slices in the x-y plane that are spaced across t according to a specified discretization $D := \{t_0 = 0, \ldots, t_K = \frac{t_{S-1}}{2}\}$ into K values. Each successive grid along t is defined by propagating along the direction of the characteristic $(x_0 + v^a(\sigma)\tau, y_0 + v^b(\sigma)\tau)$ from the previous grid, with a new column and row for $x = n \cdot t$ and $y = n \cdot t$ on the outer boundaries. Note that the first grid consists of the point $(0, 0, 0)$ only. The next step extends the characteristic out of this point and adds the correct values at the boundaries, and thus the initalization of the scheme is well defined.

To ultimately fill all grids with the correct values, we start at each grid point along the boundary and fill in values for F_σ by integrating along the characteristic direction. This fills in lines of points in the interior of the volume, and once all boundary points are processed, all interior points will be computed. For $\sigma = (k_{ab}, k_a, k_b, r)$, take a point (t_0, x_0, y_0) on the boundary, that is, $t_0 \in D$ (the discretization grid), and either $x_0 = n \cdot t_0$ or $y_0 = n \cdot t_0$. Moreover, let $F_\sigma(t_0, x_0, y_0) =: F_{\sigma 0}$. Then, for $(t_0 + \tau) \in D$, the values on the grids along the characteristic can be computed according to:

$$F_\sigma\big(t_0 + \tau, x_0 + v^a(\sigma)\tau, y_0 + v^b(\sigma)\tau\big) = e^{-H_k(\tau)}\Big(F_{\sigma 0} + \int_0^\tau g_\sigma(\alpha) e^{H_k(\alpha)} d\alpha\Big)$$

$$\text{where } g_\sigma(\tau) := \sum_{\sigma' \to \sigma} F_{\sigma'}\big(t_0 + \tau, x_0 + v^a(\sigma)\tau, y_0 + v^b(\sigma)\tau\big) Q^\rho_{\sigma',\sigma}(t_0 + \tau),$$

$$H_\sigma(\tau) := \int_0^\tau q_\sigma(\alpha) d\alpha,$$

$$q_\sigma(\tau) := -Q^\rho_{\sigma,\sigma}(t_0 + \tau), \text{ and}$$

$\sigma' \to \sigma$ is the set of states σ' that precede σ.

This scheme, which implements the method of characteristics, propagates the boundary values along the characteristic vectors. While the integral for $H_\sigma(\tau)$ is computed exactly (since we can analytically integrate $\lambda(t)$ for the spline and piecewise constant representations described in Section 3.1.1) the integral in the expression for F_σ is computed by approximating the integrand as piece-wise linear, with the pieces bridging points of the grid for σ (along a characteristic). This allows us to restrict the evaluations of g_σ, H_σ, and q_σ to points on the σ-grid.

As seen in equation (2.14), the values along the boundaries are the solution to another set of PDEs (2.16), albeit in 2 dimensions rather than 3. The boundary value PDEs can be solved using a similar scheme with the dimension reduced, following one dimensional characteristics (along which these boundary points lie). The boundary values for these secondary PDEs are $F_\sigma(t, x = y = n \cdot t) = g_\sigma^\rho(t)$ and can be computed by solving a set of ODEs using a standard numerical ODE-solver (as was done for T_{MRCA}).

We note that in solving for each state, the solutions for all states immediately upstream (states preceding σ in the ordering) are required. The grid for each state is based on the same discretization D of time-steps. However, the x and y values along the grids do depend on the characteristic direction which can differ between different states, and thus, for a given t, the x-y grid for a state σ may be different than the grid used for a preceding state σ'. Thus, in order to compute $g_\sigma(\tau)$, we obtain the values for $F_{\sigma'}$ on the x-y grid for the state σ by linearly interpolating the values from the grid of σ'. We also tested interpolation using a cubic spline, but it resulted in only marginal improvements in accuracy at the cost of significant computational time and potential numerical instabilities due to oscillations.

Efficiency of our implementation is improved by explicitly imposing the following symmetries and constraints which follow from symmetries in the ancestral process with recombination $\mathcal{A}^{\rho}(t)$:

$$F_{(k,r,r,r)}(t,x,y) = F_{(k,r,r,r)}(t,\min(x,y),\min(x,y))$$

$$F_{(k,k',k',r)}(t,x,y) = F_{k,k',k',r}(t,y,x)$$

$$F_{(k,1,0,1)}(t,x,y) = F_{k,0,1,1}(t,y,x),$$

for all $k \in \{1,\ldots,n\}$ and $k', r \in \{0,1\}$. After computing $F_\sigma(t,x,y)$ for all σ and all points on the grids, we can substitute the numerical values into (2.17) to compute the requisite probabilities for the transitions of our CHMM.

Emission probabilities are computed from the PDEs given in (2.18) using an analagous scheme for one less dimension (similar to how the boundary condition for the transition probability was solved in equation (2.16)).

2.6.4 Data Processing

Our algorithm can be applied to full genomic sequencing data for a sample of individuals from a population. Such data are often presented as genetic variation at a number of segregating sites, separated by tracts of monomorphic sites, for example, in the form of a `vcf`-file. Additionally, such datasets often have regions in the genome marked as missing. This suggests that all the sites along the genome can be grouped into tracts of two types: *non-segregating* tracts, defined as having a single segregating bi-allelic site at the beginning, followed by a stretch of non-segregating sites, and *missing* tracts, where the genetic information is flagged as missing in the data. Processing the raw data and opting to store it as a series of tracts is more efficient than storing information for each individual site, and by also storing the type of tract, the tract length, and the number of derived alleles at the head of *non-segregating* tracts, we retain all the necessary information required for our CHMM. Note that these tracts are the same monomorphic tracts which can be efficiently skipped over as described in 2.5.2. Lastly, our method needs information to distinguish between ancestral and derived alleles at each segregating site, and this is specified by providing an ancestral sequence in addition to the population sample.

2.6.5 Missing Data

Non-segregating sites in CHMM

While the *non-segregating* tracts can be handled straightforwardly by the CHMM, with the monomorphic non-segregating sites being treated as an emission of 0 derived alleles at the respective site, the missing tracts are handled slightly different. If a site is missing, the emission probabilities that are used for that site during the forward-backward algorithm are modified to reflect the fact that the missing site provides no information on the hidden state at that location. In practice, this means that each hidden state has an equal probability of emitting a missing site. For tracts of missing data, this has the effect that the posterior probability distribution across hidden states at the boundaries is informed by the surrounding sequence, but further into the interior of such tracts the posterior probability approaches the marginal distribution of states asymptotically.

Tagging

We note that genomic datasets can contain segregating sites that are not bi-allelic, for example, segregating sites with 3 alleles, or some forms of structural variation. We treat all segregating sites that are not bi-allelic as *missing* sites. This step introduces a potential bias into the CHMM because it can potentially flag individuals segregating sites as missing, while the non-segregating sites around this specific sites are still used for the analysis. A single missing site is more likely to be non-segregating a-priori, and if such a site is between two tracts of non-segregating sites, the method effectively concatenates them and interprets

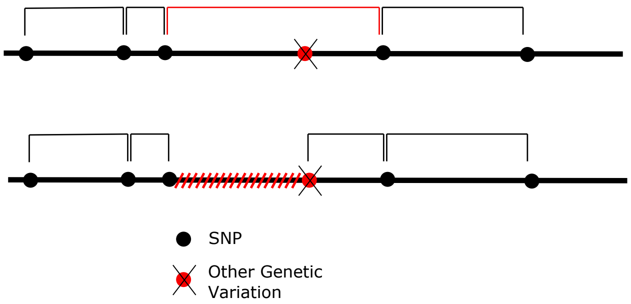


Figure 2.6: Schematic showing the improvement achieved by masking the tract upstream of a non-biallelic site as missing. In the former case (top), the tract is effectively interpreted as an artificially long region without variation (marked red), which biases the inference. By masking the left region as missing (bottom) we avoid this systematic bias.

the whole tract as an artificially long tract with no segregating sites. This biases the CHMM towards inferring shallow trees. In order to counter this systematic bias, when we tag such segregating sites that are not bi-allelic as *missing*, we also mask all the non-segregating sites upstream as missing, until we reach either a segregating site or a site marked as missing in the raw data (see Figure 2.6 for a schematic). By marking the tract upstream (or equivalently downstream) as missing as well, the true non-segregating tracts that are identified better represent the appropriate lengths of such tracts. At worst, this process loses information since the inferred trees in missing tracts approach the marginal distribution asymptotically as the algorithm moves into the interior of these tracts, but it does not introduce systematic bias.

CHAPTER 3

DEMOGRAPHIC INFERENCE WITH CHIMP

Having described the model for `CHIMP` and its implementation, we shift focus to applying this model for specific population genetic problems. Many of the implementation choices of `CHIMP` were made with its application for inferring demographic history in mind. This chapter will focus on how the model described in Chapter 2 is incorporated into a Expectation-Maximization (EM) framework that is then used to infer the population size history of a single population (from which the samples originate). We will present methodological and implementation details in Section 3.1, before benchmarking `CHIMP` against `MSMC2` and `Relate` using a series of simulation studies in Section 3.2 .

3.1 Demographic Inference Using EM

3.1.1 Parameterization of $\eta(t)$

While the population size history $\eta(t)$ is in general a non-singular positive-valued function, in order to perform efficient inference we restrict $\eta(t)$ to be parameterized by a finite number of parameters. We choose to represent $\eta(t)$ as a piece-wise constant function, where the number of pieces can be specified by the user and are uniform in $\log(t)$ space between specified bounds. In practice, it is more convenient to directly work with $\vec{\lambda}$, defined to be the piece-wise values for the coalescence rate $\lambda(t)$, and transform these back into the population size at the end.

We have also implemented the option to use a cubic-spline representation of the population size history (instead of piece-wise constant), similar to that offered by `SMC++` [16]. More specifically we represent $\lambda(t)$ as a cubic spline, and the population size history is proportional to the inverse of this. The number of nodes can be specified, and we place them equidistantly on a $\log(t)$ scale.

It is important to note that the epoch times used in parameterizing $\eta(t)$ are distinct from the $\{t_1 < \ldots < t_S\}$ used to partition T_{MRCA} (or $\mathcal{L}$) into discrete states. While other methods such as `MSMC2` choose to link these discretizations closely [15], we do not. This allows for the states to be specified entirely separately from the population parameters.

3.1.2 Expectation-Maximization Algorithm

We use the Expectation-Maximization (EM) algorithm for HHMs to infer the population size history parameters $\vec{\lambda}$. This is an iterative algorithm with each iteration consisting of two steps: an *E-step* during which each we compute the expected numbers of various transition types and various emission types, and an *M-step* where we maximize a likelihood function that is based on these expectations in order to obtain new estimates for the model parameters. Choosing initial parameters $\vec{\lambda}^0$ is described in section 3.4.1, and we denote the parameters in the k-th iteration by $\vec{\lambda}^k$.

This implementation of EM makes use of the composite likelihood scheme in a similar way to `MSMC2` [15], where the authors use all overlapping sub-groups of size $n_s = 2$ for the E-Step, and combine them in a similar way for the M-Step.

E-Step

For the k-th iteration of the E-step, we compute the initial ($\mathbf{\Pi}$), transition ($\mathbf{A}$), and emission ($\mathbf{B}$) probabilities under the coalescent rate function given by $\vec{\lambda}^k$, and use these probabilities to perform the Forward-Backwards algorithm across the emission sequence $\vec{d}$ (all using the machinery described in chapter 2).

The Forward-Backward algorithm yields the likelihood of observing the data under the current demographic parameters and the inferred hidden state sequence, and, consequently, it also gives, conditional on the current parameters and the data,

$$\mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \to j) \text{ transitions}]$$

$$\mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \downarrow d) \text{ emissions}]$$

$$\mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[s^1 = i],$$

that is, the expected number of transitions from state i to j, the expected number of times state i emits d derived alleles, and the expected occupation of an initial state, respectively. These expectations, conditional on the observed data in our composite likelihood model (section 2.5.4), can simply be computed by adding up the expectations for each sub-sample of n_s haplotypes of genomic data . Thus the run-time of the E-step (as we have implemented it, without parallelization) scales linearly with the number of total samples, linearly with the sum of chromosome lengths being analyzed, and quadratically with the number of hidden states.

M-step

After each E-step, we perform an M-step during which we update the values of $\vec{\lambda}$ by numerically maximizing the objective function, defined as the expected log-likelihood of $\vec{\lambda}$, with respect to the conditional distribution of the hidden states given the data and the current parameter estimates $\vec{\lambda}^k$,

$$\begin{aligned} Q(\vec{\lambda}|\vec{d},\vec{\lambda}^k) = &\sum_{i\in\mathcal{S}} \log\left(\Pi_i(\vec{\lambda})\right) \cdot \mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[s^1 = i] \\ &+ \sum_{i,j\in\mathcal{S}} \log\left(A_{ij}(\vec{\lambda})\right) \cdot \mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \to j) \text{ transitions}] \\ &+ \sum_{i\in\mathcal{S},d\in\mathcal{D}} \log\left(B_{id}(\vec{\lambda})\right) \cdot \mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \downarrow d) \text{ emissions}], \end{aligned} \tag{3.1}$$

where we explicitly denote the initial, transition, and emission probabilities as functions of $\vec{\lambda}$ to stress that they are computed for the parameters that we optimize over. The parameters that maximize this function yield the updated parameters for the next iteration and are given by

$$\vec{\lambda}^{k+1} := \underset{\vec{\lambda}}{\operatorname{argmax}} \left[Q(\vec{\lambda} | \vec{d}, \vec{\lambda}^{k}) \right].$$

This objective function can be modified to artificially penalize population size histories that don't conform to a priori notions about the form of natural histories. In Section 3.4.2, we provide details on our implementation of a set of regularizing coefficients that allow the user to artificially enforce varying degrees of smoothness on the results of inference. While this option has been implemented, no regularization was used for the simulation studies presented in Section 3.2.

Since `CHIMP` evaluates the Q function numerically, we use the Nelder-Mead simplex optimization procedure for numerical optimization [32]. Section 3.4.3 includes some details about this optimization scheme, and some of the hurdles we encountered. The optimization is performed in a search space of logarithmic coalescence rates, which is a uniquely robust space in which to perform optimization of coalescent rates [33] and also has the benefit that the coalescent rates are positive by design. After finding the optimal coalescent rates using the EM algorithm, we invert and scale them to recover the estimates for the population size history $N(k)$. Regarding efficiency, since the expectations in the objective function are summed over all the sub-groups of samples and chromosomes provided for the analysis, the run-time of the M-step is constant for different sample sizes and total chromosomal lengths.

3.2 Simulation Studies

To evaluate the accuracy of our method, we performed a series of simulation studies. We generated data under various demographic scenarios, and then inferred the population size history from these simulated datasets using CHIMP-T_{MRCA} and CHIMP-$\mathcal{L}$, with T_{MRCA} and $\mathcal{L}$ as the hidden state, respectively, and compared the results to inference using MSMC2 [15, v2.1.2] and Relate [10, v1.1.3]. For each study we used the specified model of the demographic history to simulate $m = 16$ replicates of data using msprime [25, v0.7.4], where each replicate consists of $n = 200$ haplotypes of length 200 Mbp. The per generation per site recombination and mutation rates we used were $r = \mu = 1.25 \cdot 10^{-8}$, to mirror applications to human genetic data. We inferred the population size history for each of the replicates using the different methods and visualized the variability of the estimates across replicates.

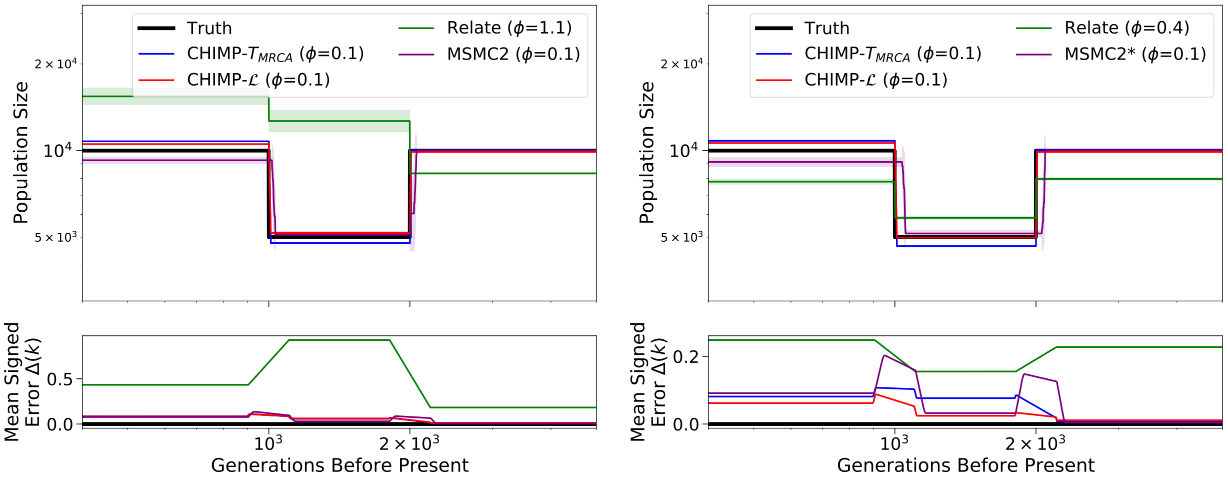
We noticed that the performance of Relate improved when we simulated and analyzed data with a human recombination map (see Section 3.4.5 and Figure 3.8). This is likely due to the fact that Relate benefits from *cold* spots (regions of low recombination rate) in the recombination map. However, the performance of CHIMP and MSMC2 were not substantially adversely affected when they were run with the (inaccurate) assumption of a constant recombination rate. For this reason, we conducted our simulation studies with a uniform recombination map.

To demonstrate the performance of the different methods for different samples sizes, we present two analyses for each demographic model, one using the full 200 haplotypes simulated, and one using a subset of 10 haplotypes chosen uniformly at random. For CHIMP, we ran the analyses using the full EM algorithm for $n = 10$. For $n = 200$, we used the composite likelihood framework detailed in Section 2.5.4, with sub-groups of size $n_s = 10$, a value that we found performs well in terms of efficiency and accuracy. The method MSMC2 can analyze all pairs of haplotypes in the dataset, which we did for the samples of size $n = 10$, but was computationally prohibitive for samples of size $n = 200$. For the later case, we instead restricted MSMC2 to analyze a subset of 50 non-overlapping pairs of haplotypes (100 haplotypes total) since memory requirements of the method became a limitation beyond this number. Moreover, in an effort to ensure a fair comparison between the methods, we ran all analyses for a piecewise constant parametrization of the population size history, and chose the same change points across the methods. The change points could be explicitly specified for CHIMP and Relate. For MSMC2, specifying change points was achieved by providing a time-segment pattern (as required by the method) that placed the change points as close to the desired ones as possible. This yielded a close match in most cases, with minor inaccuracies in more recent times and very ancient times. To initialize the iterative inference methods, we chose Watterson's estimator for CHIMP, as detailed in Section 3.4.1, and the default initialization for MSMC2.

To aid visualizing and summarizing the performance of a method in a specific setting, as well as comparing the results between methods, we also plot the mean absolute deviation from the true population size across the replicates in generation k

$$\Delta(k) := \frac{1}{m} \sum_{j=1}^{m} \left| \log \left(\frac{\hat{N}^{(j)}(k)}{N_{\text{true}}(k)} \right) \right|,$$

where $\hat{N}^{(j)}$ is the population size estimated in replicate j, and for each method, compute the integral of this quantity $\phi = \int_{k_{\text{min}}}^{k_{\text{max}}} \Delta(k) \frac{1}{k} \, dk$ as a measure of discrepancy from the truth over the full history. Here k_{min} and k_{max} are the minimum and maximum of the respective discretizations with one logarithmic discretization step subtracted and added, respectively. Note that the factor $\frac{1}{k}$ suppresses deviations in the distant past and transforms the integral to a regular integral of Δ on a $\log(k)$ timescale, which matches the visualization more closely.



(a) Sample size 10.

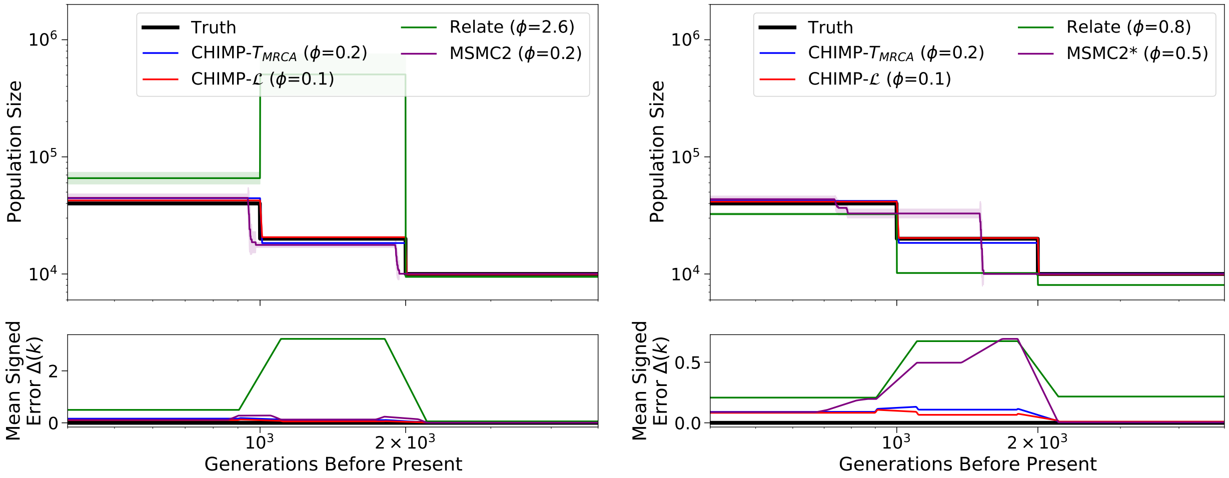
(b) Sample size 200.

Figure 3.1: Results of inference in the bottleneck scenario for sample size (a) 10 and (b) 200. We compare the results of `CHIMP`, `MSMC2`, and `Relate` to infer the three population sizes, fixing the change points to match the truth (shown in black). Solid lines are averages over 16 replicates and shaded area indicates standard deviation. Mean signed error $\Delta(k)$ is shown at bottom and has been smoothed using moving average for visualization purposes. The integral ϕ is indicated in the legend. (*) Note that for sample size 200, `MSMC2` was only run on 50 non-overlapping pairs.

3.2.1 Demographic Models with Limited Number of Parameters

We first compared the methods on two simple demographic models with three parameters to estimate. The first models a population that experienced a bottleneck event. In this scenario, the ancestral population is of constant diploid size 10,000. At 2,000 generations before present, the population size is reduced to 5,000, but recovers to 10,000 at 1,000 generations before present. The second scenario models a population experiencing piecewise growth. Here, the ancestral population size is again 10,000 diploid individuals. At 2,000 generations before present, the population size doubles to 20,000, and then doubles again at 1,000 before present to 40,000. In both scenarios, we fixed the demographic parameterization for each method to estimate a three-parameter piecewise constant population size history with change points matching those of the true population size histories.

Figure 3.1 shows the results of the inference using the different methods in the bottleneck



(a) Sample size 10.

(b) Sample size 200.

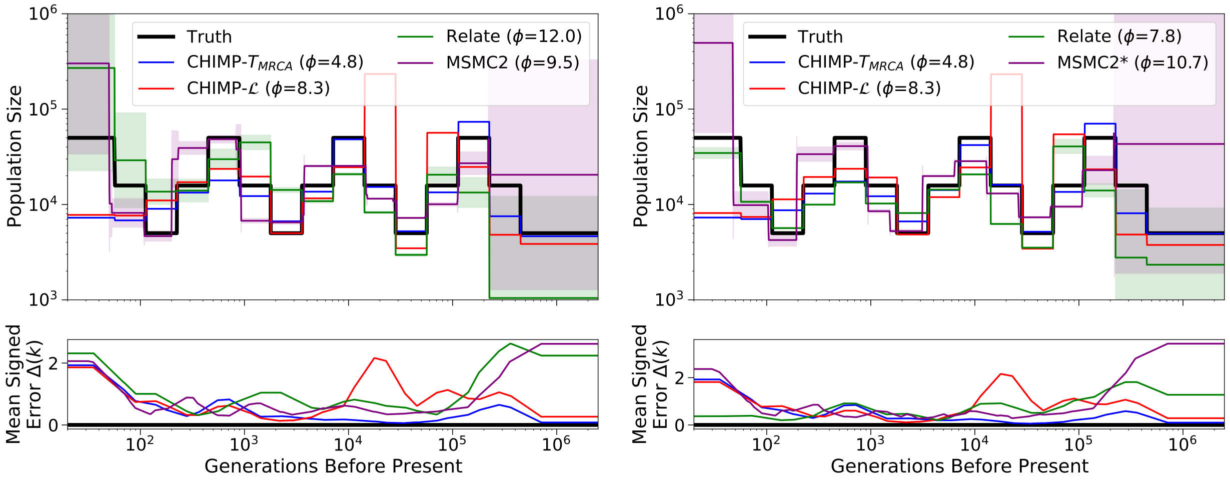
Figure 3.2: Results of inference in the piecewise growth scenario for sample size (a) 10 and (b) 200. We compare the results of `CHIMP`, `MSMC2`, and `Relate` to infer the 3 population sizes, fixing the change points to match the truth (shown in black). Solid lines are averages over 16 replicates and shaded area indicates standard deviation. Mean signed error $\Delta(k)$ is shown in bottom plot and has been smoothed using moving average for visualization purposes. The integral ϕ is indicated in the legend. (*) Note that for sample size 200, `MSMC2` was only run on 50 non-overlapping pairs.

scenario for sample sizes 10 and 200. In this case, CHIMP-T_{MRCA}, CHIMP-$\mathcal{L}$, and MSMC2 recover the true population size history accurately and show little variability across replicates. Relate, however, does not recover the underlying true size accurately. For a sample of size 10, it does not infer a bottleneck, and while for sample size 200, a bottleneck is inferred, the ancestral and contemporary population size are not inferred correctly. Nonetheless, the inference shows little variability.

The results for the inference in the piecewise growth scenario are depicted in Figure 3.2. In this scenario, CHIMP-T_{MRCA} and CHIMP-$\mathcal{L}$ are again able to infer the population size history with high accuracy. For a sample of size 10, MSMC2 also recovers the true size, however, for samples size 200, the inference for the intermediate size is systematically biased, with little variability. The reason for this deviation is likely the fact that when using MSMSC2 for large sample sizes, the interface does not allow setting the change points for the inference closer to the true values. Given this constraint, the inferred population sizes are quite accurate and we believe that the method would have high accuracy if the appropriate boundaries could be specified. Again, for a small sample, Relate overestimates especially the intermediate size, and for the large sample size of 200 the estimates are not as accurate as MSMC2 and CHIMP.

3.2.2 Inference for Piecewise Constant Sawtooth Demography

A benchmark for population size inference that has been used in recent studies is a population size history that exhibits oscillations, referred to as *sawtooth* history [14, 16, 10, 20]. We will analyze a continuous version of this scenario in Section 3.2.3, but we were first interested in comparing the performance of the methods for a piecewise constant version so that the true population size history could in principle be exactly recovered using the different methods. To this end, we simulated data under a piecewise constant sawtooth history, were the population size oscillates between 50,000, 15,811, and 5,000 at 14 change points that are equidistant on a logarithmic scale between 57 and 448,806 generations before present. Again, we simulated



(a) Sample size 10.

(b) Sample size 200.

Figure 3.3: Results of inference in the piecewise sawtooth scenario for sample size (a) 10 and (b) 200. We compare the results using `CHIMP`, `MSMC2`, and `Relate` to infer the population sizes in the intervals, fixing the change points to match the truth (shown in black). Solid lines are averages over 16 replicates and the standard deviation is indicated by the shaded areas. Mean signed error $\Delta(k)$ is shown in bottom plot and has been smoothed using moving average for visualization purposes. The integral ϕ is indicated in the legend. Note that `MSMC2` groups epochs in the very distant past due to limits of the method interface. (*) For sample size 200, `MSMC2` was run on 50 non-overlapping pairs, and only 15 of the 16 replicates completed successfully.

16 replicates for this scenario and inferred the population sizes using the different methods for different sample sizes, while keeping the change points fixed to the true values.

The results of this simulation study are depicted in Figure 3.3. We observe that CHIMP-T_{MRCA} estimates the population sizes in the respective intervals accurately in the intervals 500 generations before present and further in the past. The estimates are accurate even in the very distant past. It smooths the history in the very recent intervals. The accuracy does not change substantially when using samples of different sizes. In general, CHIMP-$\mathcal{L}$ behaves more erratically. It also does not infer the very recent times correctly, and is only correct for some of the intermediate intervals. This is likely due to the fact that we infer many demographic parameters (when compared to the inference in Section 3.2.1) which results in a higher dimensional inference problem with a likelihood surface that is more difficult to navigate and causes the method to converge to a local optimum. The fact that the direction of the bias replicates over different datasets suggests that the initial parameter choice and the details of the numerical optimization procedure (Nelder-Mead algorithm) affect the navigation to the local optima.

MSMC2 shows accurate performance for intermediate times despite smoothing out some of the oscillations, but demonstrates high variability and a systematic upward bias below 100 generations and above 100,000 generations before present. Its accuracy does not change much between analyzing samples of different sizes. Relate has a high variability and upward bias for very recent times if the sample size is low, but the recent sizes are very accurately estimated when using a large sample size. The accuracy for intermediate and ancient times is not very high, and this performance is only slightly improved in intermediate times for larger sample sizes. Ultimately, between the different methods tested, each performs better in some timeframe or for specific sample sizes and worse for others. Measured in terms of integrated mean signed error ϕ, CHIMP-T_{MRCA} shows the overall best performance in this demographic scenario.

3.2.3 Inference for Continuously Varying Population Size History

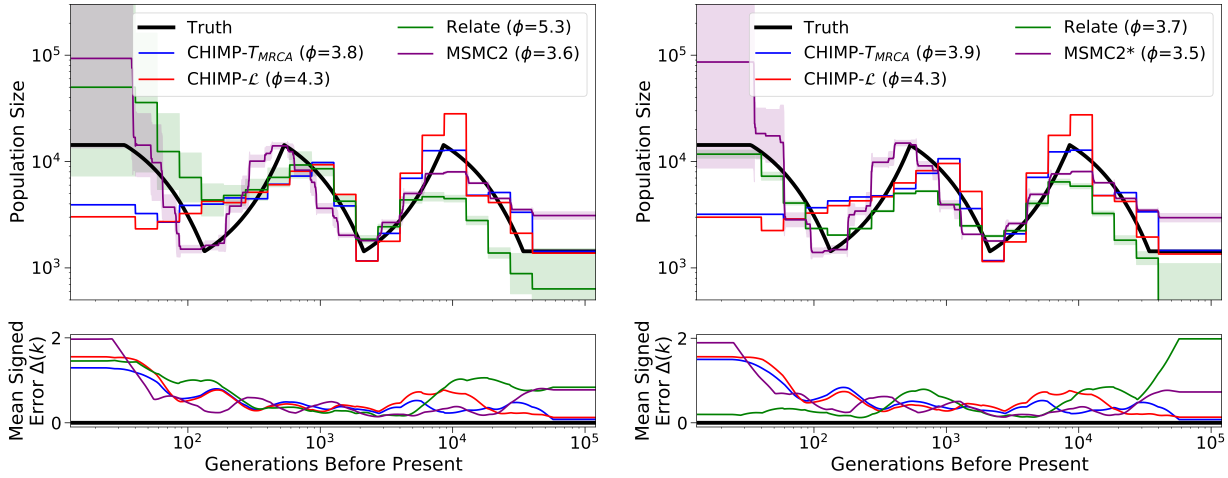
Lastly, we we studied the performance of the inference methods on models of continuously varying population size history. Specifically, we considered the (continuous) *sawtooth* model implemented in `stdpopsim` [34, *ID=Zigzag_1S14*]. In this model, the population size alternates between a maximum of 14,312 and a minimum of 1,431, with three maxima and three minima equidistant on a logarithmic scale between 33 and 34,133 generations before present. Note that the maxima and minima are roughly a fifth of the ones used by [14] and [16]. We nonetheless decided to use this model here to investigate performance over a wider range of parameters in our simulation study. The second model we considered here is a bottleneck followed by exponential growth, a cartoon of an Out-Of-Africa population size history [35, 36]. In this model, the ancestral population size of 10,000 sharply drops to 2,000 at 4,000 generations before present. At 1,000 generations before present, the population size starts growing up to the present at an exponential rate of 0.25% per generation. Again, we simulated 16 replicates in each scenario with 200 haplotypes of length 200 Mbp and analyzed each replicate with each method on the full sample and on a subsample of size 10. We used the same discretization across methods for a better comparison, first specifying a minimum and maximum time and then choose 19 equidistant change points between these values (inclusive) on a logarithmic scale. The minimum and the maximum time were 40 and 40,000 for the sawtooth, and 200 and 20,000 for the bottleneck followed by growth scenarios, respectively.

The results in the sawtooth scenario are shown in Figure 3.4. We observe that for `CHIMP`-T_{MRCA}, `CHIMP`-$\mathcal{L}$, and `MSMC2`, the accuracy does not differ substantially between the different sample sizes. Again, `CHIMP`-T_{MRCA} smooths the population sizes earlier then 500 generations before present, thus underestimating the recent size, and smoothing out part of the first peak. The more ancient part of the first peak, the second peak and the ancestral population size are captured accurately. `CHIMP`-$\mathcal{L}$ performs similarly for recent and intermediate times, but behaves erratically around 10,000 generations before present. Again, we suspect that this is due to the difficulty in navigating the high dimensional likelihood surface. MSMC2 captures both peaks, but slightly overestimates the ancestral size and substantially overestimates the recent sizes with a high degree of variability between replicates. For a small sample size, Relate overestimates recent sizes. It does infer two peaks, but the sizes and timing do not fully align with the truth. For a large sample size, Relate infers recent population sizes with high accuracy, but still underestimates the sizes of the two peaks. In terms of integrated mean signed error ϕ summarizing the overall accuracy, CHIMP-T_{MRCA} and MSMC2 perform comparably for small sample sizes, and for large samples, Relate reaches the same level of accuracy.

Figure 3.5 shows the results in the scenario where a bottleneck is followed by exponential growth. In this scenario, all four methods capture the general trend of the population size history. Again, the performance of CHIMP-T_{MRCA}, CHIMP-$\mathcal{L}$, and MSMC2 does not differ substantially between sample sizes used in the analysis, however, Relate overestimates the size history when using a small sample, but underestimates the history when using a large sample. MSMC2 and Relate smooth out the abrupt decline of the population size at the beginning of the bottleneck, whereas CHIMP-T_{MRCA} and CHIMP-$\mathcal{L}$ do infer a sharper decline. However, the latter methods show inaccuracies at the more ancient times, which are more pronounced for CHIMP-$\mathcal{L}$. In this scenario, the method MSMC2 exhibits the best overall performance metric ϕ.

3.2.4 Computational Efficiency

In general, the runtime of the CHMM methods CHIMP and MSMC2 scale linearly with the number of loci L. In addition, the CHIMP methods scale linearly with the sample size (due to the composite likelihood framework introduced in Section 2.5.4). If all pairs of samples are used in MSMC2, the method scales quadratically with the sample size, but when using



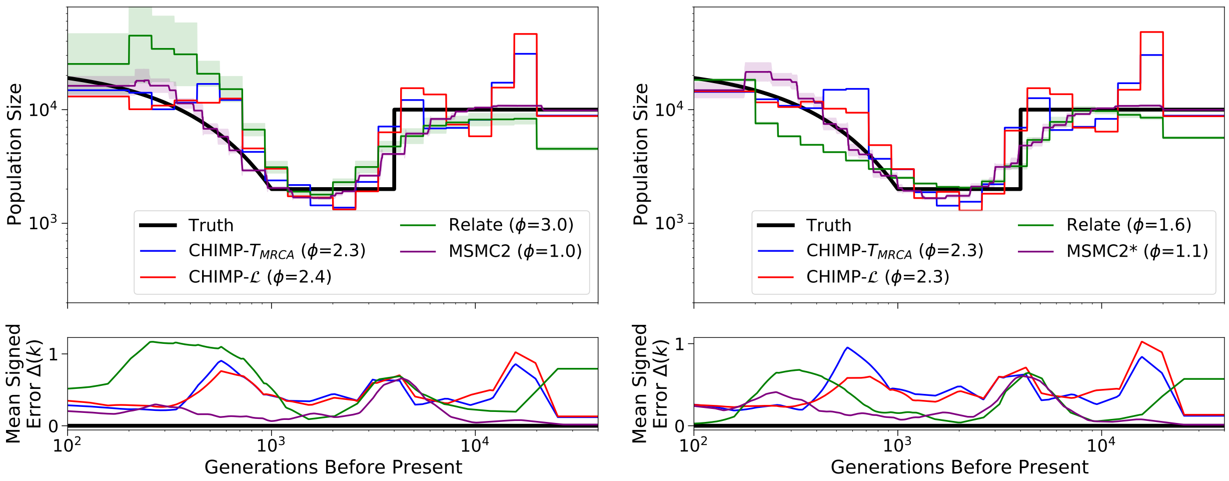
(a) Sample size 10.

(b) Sample size 200.

Figure 3.4: Results of inference in the continuous *sawtooth* scenario for sample size (a) 10 and (b) 200. We compare the results of `CHIMP`, `MSMC2`, and `Relate` using a piecewise constant population size history with 19 change points. Truth shown in black. Solid lines are averages over 16 replicates and shaded areas indicate standard deviation. Mean signed error $\Delta(k)$ is shown at bottom and has been smoothed using moving average for visualization purposes. The integral ϕ is indicated in the legend. (*) For sample size 200, `MSMC2` was run on 50 non-overlapping pairs.

	CHIMP-T_{MRCA}	CHIMP-$\mathcal{L}$	MSMC2*	Relate
Bottleneck ($n = 10$, Fig. 3.1a)	0.1	1.3	4.9	0.1
Bottleneck ($n = 200$, Fig. 3.1b)	1.2	2.4	5.4	7.0
Piecewise Growth ($n = 10$, Fig. 3.2a)	0.1	1.7	2.1	0.1
Piecewise Growth ($n = 200$, Fig. 3.2b)	1.1	2.9	9.7	9.2
Piecewise Sawtooth ($n = 10$, Fig. 3.3a)	0.9	29.7	4.3	0.2
Piecewise Sawtooth ($n = 200$, Fig. 3.3b)	7.2	35.2	6.2	8.9
Sawtooth ($n = 10$, Fig. 3.4a)	1.2	29.9	4.1	0.1
Sawtooth ($n = 200$, Fig. 3.4b)	7.6	37.9	**5.0	3.0
Bottle + Growth ($n = 10$, Fig. 3.5a)	1.2	28.3	5.7	0.1
Bottle + Growth ($n = 200$, Fig. 3.5b)	7.5	35.1	3.2	3.7

Table 3.1: Run-times in hours for the analysis of simulated data in the different scenarios, averaged over the respective 16 replicates in each case. The runtimes for `MSMC2` are slightly inflated, as the number of CHMM states had to be increased to allow for the closest matching of demographic epochs. (*) For the $n = 200$ scenarios, `MSMC2` was only run on 50 non-overlapping pairs of samples. (**) For this set of inferences, only the 15 out of the 16 replicates that successfully terminated were averaged.



(a) Sample size 10.

(b) Sample size 200.

Figure 3.5: Results of inference in the bottleneck followed by growth scenario for sample size (a) 10 and (b) 200. We compare the inference of CHIMP, MSMC2, and Relate using a piecewise constant population size history with 19 change points. Truth shown in black. Solid lines are averages over 16 replicates and shaded areas indicate standard deviation. Mean signed error $\Delta(k)$ is shown in bottom plot and has been smoothed using moving average for visualization purposes, The integral ϕ is indicated in the legend. (*) For sample size 200, MSMC2 was run on 50 non-overlapping pairs.

non-overlapping pairs, like in our analysis of large samples, it scales linearly. Relate scales linearly with number of loci and quadratically with samples size. However, the method is implemented very efficiently and allows fast reconstruction of multi-locus genealogies for large sample sizes.

To exhibit the actual computational performance of the different methods in the simulation study, we list the average run-times in the different scenarios in Table 3.1. Since we ran MSMC2 using 50 non-overlapping pairs of samples instead of the full 200, these are the times that we report. Additionally, runtimes for MSMC2 are slightly inflated, as the number of CHMM states had to be increased to allow for the closest matching of demographic epochs. For both CHIMP methods, we observe a difference in performance between small and large samples, and a difference between the scenarios with few parameters to infer and the more general scenarios. The latter is a result of fast convergence to an optimum when only a few parameters describe the model, whereas convergence is slower in a higher dimensional parameter space. The performance of MSMC2 shows little variability across sample size and scenarios. Since for a sample of size $n = 10$, we analyze all $\binom{10}{2} = 45$ pairs, it is expected that the performance is similar to analyzing 50 non-overlapping pairs. The performance of Relate depends on the sample size, but shows little variability across scenarios, as expected, since the reconstruction of the genealogy is not strongly affected by the parameterization of the demographic model.

For sample size $n = 10$ and a small number of parameters, inference with CHIMP-T_{MRCA} is comparably fast to Relate, the former slowing down when the number of parameter increases, but both are faster than MSMC2. For $n = 200$, CHIMP-T_{MRCA} performs comparably to the other methods in most scenarios. The method CHIMP-$\mathcal{L}$ is substantially slower than the other methods, especially in the scenarios where the demographic model is described by many parameters. In these scenarios, the EM algorithm requires many steps to converge, and each step requires evaluating PDEs to compute the transition and emission probabilities.

This is computationally more expensive than, for example, evaluating ODEs as required for CHIMP-T_{MRCA}.

3.2.5 Analyzing Unphased and Pseudo-haploid Data

Our method CHIMP can also be readily applied to unphased genomic data, and is therefore a promising method for applications where high quality phased genomes are not available, for example in human ancient DNA or for non-model organisms. In the simulation studies presented in this paper, we used simulated data, which is perfectly phased. However, the parameter inference performed using CHIMP only uses the number of derived alleles at each locus as input, and is therefore invariant to any phasing of the data. Thus, inference under the method will not be affected by phasing errors, and can even be performed on completely unphased data. MSMC2, when run on all possible of pairs of samples, requires phased data. If it is run on non-overlapping pairs of haplotypes where each pair is associated with a single individual, as we did here for large samples, it could be run on unphased data. However, such a scheme is not commonly used in the literature. Since Relate reconstructs multi-locus genealogies relating haplotypes, it cannot be applied to unphased data and will be adversely affected by phasing errors.

In addition to being able to analyze unphased data, our method can also take a form of pseudo-haploid data as input. Generating pseudo-haploid data is a strategy often applied to low-coverage sequencing data, where reliable diploid genotype calls are not feasible, and may introduce unwanted biases, for example for ancient human DNA [37]. In pseudo-haploid data, at each SNP, one sequencing read covering the respective SNP is chosen uniformly at random, and the allele on this read is then reported as the haploid genotype for the individual. We implemented an option for CHIMP that extends the CHMM to pseudo-haploid data. To analyze pseudo-haploid data for a sample of size n, we implement a two layered emission model. The CHMM is implemented using the T_{MRCA} for a sample of size $2n$ as the hidden

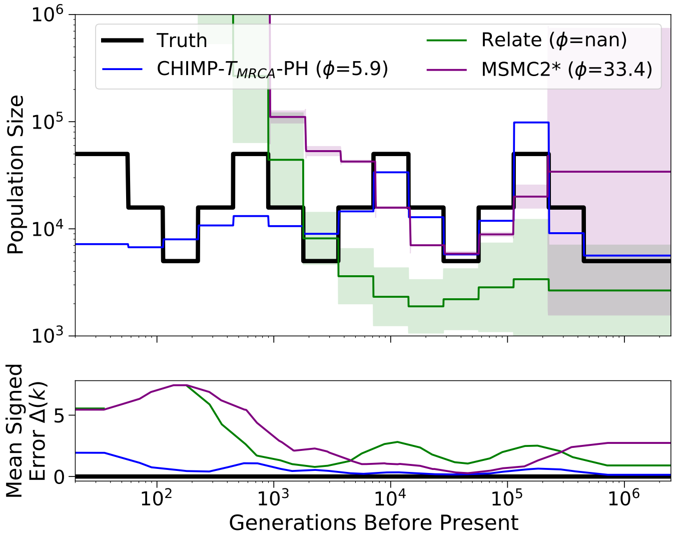


Figure 3.6: Results of inference using 10 pseudo-haploids simulated under the piecewise sawtooth demography. We compare the results of $\texttt{CHIMP} - T_{MRCA}$ using the pseudo-haploid option, `MSMC2`, and `Relate`, fixing the change points to match the truth (shown in black). Solid lines are averages over 16 replicates and shaded area indicates standard deviation. Mean signed error $\Delta(k)$ is shown in bottom plot and has been smoothed using moving average for visualization purposes. The integral ϕ is indicated in the legend. Note that `Relate` entirely failed to estimate a population size in the most recent epochs, resulting in indeterminate values. (*) For `MSMC2`, only 12 of the 16 runs completed successfully.

state with the respective transition probability. At each locus, the emission in the first layer is then the number of derived alleles in a sample of size $2n$. In the second layer, this sample is then down-sampled to a number of derived alleles in a sample of size n using hypergeometric probabilities.

We performed an additional simulation study, to demonstrate the inference from pseudo-haploid data using our method. To this end, we simulated 20 haplotypes of length 200 Mbp using the piecewise constant sawtooth demography (see Section 3.2.2). For each pair of haplotypes (diploid individual), at each locus, we selected one of the two alleles uniformly at random to obtain a dataset of 10 pseudo-haploid samples. We performed inference on the 10 pseudo-haplotypes of this data using `CHIMP`-T_{MRCA} with the pseudo-haploid option. We compared the results to those obtained using `MSMC2` and `Relate` for which we naively treated the data as if it were diploid data in order to perform demographic inference. The results are shown in Figure 3.3a. Not surprisingly, `MSMC2` and `Relate` do not infer the correct population size history, with particularly large errors in recent times, whereas `CHIMP` retains an accuracy close that demonstrated for full diploid data.

While these results look promising, it is important to note that `CHIMP`'s performance makes use of the information that there are no segregating sites between the SNPs in the data, which might not be a valid assumption in low coverage sequencing data. Despite this, we do believe that the capability to analyze pseudo-haploid data generated in this way presents an exciting avenue for future extensions.

3.3 Conclusions

Table 3.2 shows a summary of the performance and some features of the different methods that we compared in our simulation study. `CHIMP`-T_{MRCA}, `CHIMP`-$\mathcal{L}$, and `Relate` can be applied to samples of arbitrary sizes, whereas `MSMC2` is limited in this regard. Furthermore, `CHIMP`-T_{MRCA}, `CHIMP`-$\mathcal{L}$ can be applied to unphased data, and in limited capacity to pseudo-

	Gener. bp	CHIMP-T_M	CHIMP-$\mathcal{L}$	MSMC2	Relate
Sample size		Large	Large	Limited	Large
Parametrization		Flexible	Flexible	Limited	Flexible
Accuracy	Few param.	High	High	High	Low
	$k < 100$	Low	Low	Low	High (large n)
	$100 < k < 500$	Low	Low	High	High (large n)
	$500 < k < 10^5$	High	Mixed	High	Mixed
	$k > 10^5$	High	High	Low	Low
Runtime		Fast	Slow	Fast (low n)	Fast
Unphased data		Yes	Yes	Possible	No
Pseudo-haploid		Limited	Limited	No	No

Table 3.2: Summary of the performance and features of the different methods compared in our simulation study. The ranges are given in generations before present.

haploid data. but Relate requires phased data. MSMC2 can be applied to unphased data, if the appropriate pairs of haplotypes are chosen for the analysis. CHIMP-T_{MRCA} and Relate can be used to analyze large samples quickly, whereas CHIMP-$\mathcal{L}$ was very slow in comparison. The runtime of MSMC2 was comparable with CHIMP-T_{MRCA} and Relate, but the method could not be run on the full sample of size $n = 200$. Moreover, CHIMP-T_{MRCA}, CHIMP-$\mathcal{L}$, and Relate are very flexible in terms of user-specification of the demographic model, whereas MSMC2 limits the user to choose an appropriate time-segment string.

The inference using CHIMP-T_{MRCA}, CHIMP-$\mathcal{L}$, and MSMC2 showed very high accuracy when analyzing scenarios with a limited number of demographic parameters. In contrast, Relate did not perform well in this case. When performing inference under a flexible piecewise constant parametrization, CHIMP-T_{MRCA} has limited accuracy in recent and intermediate times. CHIMP-$\mathcal{L}$ did perform worse in intermediate times, but both methods demonstrated high accuracy in ancient times. MSMC2 did not infer the population size history well in recent and ancient times, but showed the best performance among the methods tested here in intermediate times. Relate inferred recent population sizes well, if a large sample was used, but the performance was less accurate for small samples and intermediate to ancient times.

We note that the exact time-frames depend on the baseline effective population sizes. The scenarios that we investigated here ranged from $N_e \approx 4,000$ to $12,000$.

When conducting the analyses presented here and designing comparisons between the methods that can be deemed fair, the flexibility of the user-interface and the heuristics used in choosing an a priori discretization of the population size history impacts the applicability and performance of the different methods. We demonstrate this in Figure 3.7 where we infer the size history of data simulated under the bottleneck followed by growth scenario, where all methods are run using their default parameters. In Section 3.4.4 we discuss the heuristic `CHIMP` uses for determining the default parameterization, and we find it to be robust in the scenarios that we considered and believe that it performs well in general. However, exploring optimal ways of parameterizing models with no prior information about the demographic history is an important area in which further study is needed. In this context, parameter free approaches, see for example work by [38], present interesting alternatives.

Overall, `CHIMP`-T_{MRCA} performs comparably to the other methods tested here in most scenarios when inferring sizes beyond 500 generations before present, and runs quickly on large datasets. An advantage is the fact that it can be applied to unphased and, in limited capacity, pseudo-haploid data; thus it offers a useful alternative to other existing methods in situations where high quality data is not available. Overall, the inference accuracy and run-time of `CHIMP`-$\mathcal{L}$ was very poor. We thus do not recommend this approach for inference of populations size histories, unless improvements can be made in terms of efficiently computing the probabilities required for the CHMM and navigating the high dimensional optimization problem. `MSMC2` showed very high inference accuracy for intermediate times and thus proves to be an effective method if the sample size is not too large. `Relate` is fast and can be applied to large samples. The inference accuracy is good with sufficient samples, especially in recent times, but suffers if the recombination map does not have *cold* spots.

3.4 Supplementary Details

3.4.1 Initialization of Population Size History

While we allow for the user to directly specify a population size history which is used to initialize the EM algorith, the default initialization chooses a constant population over time ($N(k) = \hat{N}$). For this constant value $\hat{N}$, we use the effective population size inferred from the data using Watterson's estimator of $4\hat{N}\mu$, fixing the mutation probability μ to a value appropriate for the data analyzed. We also use the value for $\hat{N}$ to guide choosing a partition for the CHMM states (Section 2.6.1) and in computing regularizing coefficients (Section 3.4.2).

3.4.2 Regularization of Inference Results

For demographic inference problems, we might have some prior notion of the shape of the population size history function. For different applications we may have different tolerances for sharp changes in population size, or may expect varying strengths of deviations from flat histories.

Including regularization parameters helps to suitably constrain the inferred population size histories. For a given choice of population size history parameters that define a coalescence rate function $\lambda(t)$, we define the following four regularization quantities, where $\hat{\lambda}$ is the coalescent rate corresponding to Watterson's estimate for the effective population size in a constant model, computed from the data:

$$\begin{aligned}
R_{02}(\lambda(t)) &= \int \left(\lambda(t) - \hat{\lambda}\right)^2 dt \\
R_{11}(\lambda(t)) &= \int |\lambda'(t)| dt \\
R_{12}(\lambda(t)) &= \int \left(\lambda'(t)\right)^2 dt \\
R_{22}(\lambda(t)) &= \int \left(\lambda''(t)\right)^2 dt.
\end{aligned}$$

The quantity R_{22} is analogous to the regularization quantity that `SMC++` uses, if specified [16].

We enforce these regularization constraints into our method by modifying the objective function during the M-step. The parameter update during EM (from (3.1)) then becomes

$$\vec{\lambda}^{k+1} = \text{argmax}_{\vec{\lambda}} \left[Q(\vec{\lambda}|\vec{d}, \vec{\lambda}^k) - \sum_i c_i R_i \right],$$

where i spans the various regularization types and c_i are the respective weights used to adjust their relative strengths. The user can forgo the option to include any regularization by simply selecting 0 for the coefficients (the default behavior of our software implementation).

We note that the above definitions of the regularization quantities are can be readily computed when the coalescent rate functions are continuous (as in the spline representation), however for the default piece-wise constant representation of $\lambda(t)$ we use the generalized notion of derivatives and integrals for discrete domains instead. The derivatives are computed from finite differences between the constant values, and integration becomes summation over the constant pieces.

We also include an option to compute the regularization quantities in $\log(t)$ space instead. The reason for this additional option is that population size histories are often visualized in this logarithmic scale.

3.4.3 Simplex-Based Optimization

For the M-step of our inference method we opted to use the Nelder-Mead simplex optimizer as implemented in the Apache Commons Library (https://commons.apache.org/, v3.6). Simplex optimizers do not require an analytic objective function and use relatively low numbers of function evaluations. This is advantageous for our purpose since our objective function is not easily differentiable and is also computationally intensive to compute, since it involves numerically solving ODE or PDE systems. We also tried the Apache Commons library's implementation of a Powell optimizer, however, for our test scenarios this yielded very inaccurate results compared to the Nelder-Mead optimizer.

Even when using the Nelder-Mead method, we found that inference results were highly dependent on the initial setup for the simplex. At the start of each M-step, a simplex is initialized based on the current best estimate of the demographic parameters $\vec{\lambda}^k$. During out explorations, we found that the inference results depend on the shape and orientation of this simplex. Among the simplices we compared are: **(1)** Apache Commons' default implementation, where $\vec{\lambda}^k$ was used as the first point of the simplex, and all subsequent points were determined by taking the successive steps $(1, 0, 0, ...0)$, $(1, 1, 0, ...0)$, $(1, 1, 1, ...0)$,...,$(1, 1, 1, ...1)$ **(2)** an alternating version of the default where the simplex points are determined by taking the steps $(1, 0, 0, 0...)$, $(1, -1, 0, 0, ...)$, $(1, -1, 1, 0, ...)$,$(1, -1, 1, -1, ...)$ instead (to mitigate some of the directional bias) **(3)** An equilateral simplex, formed by the steps $(1, 0, 0, ...)$,$(0, 1, 0, ...)$, $(0, 0, 1, ...)$..., and the equidistant point in the $(-1, -1, -1, ...)$ direction, which is then translated to be centered on $\vec{\lambda}^k$ (it isn't used as a vertex) **(4)** An equilateral simplex centered on $\vec{\lambda}^k$ which is then given a random n-dimensional rotation. We found that the equilateral simplex that was consistently oriented as described in option **(3)** yielded the most numerically stable and accurate results.

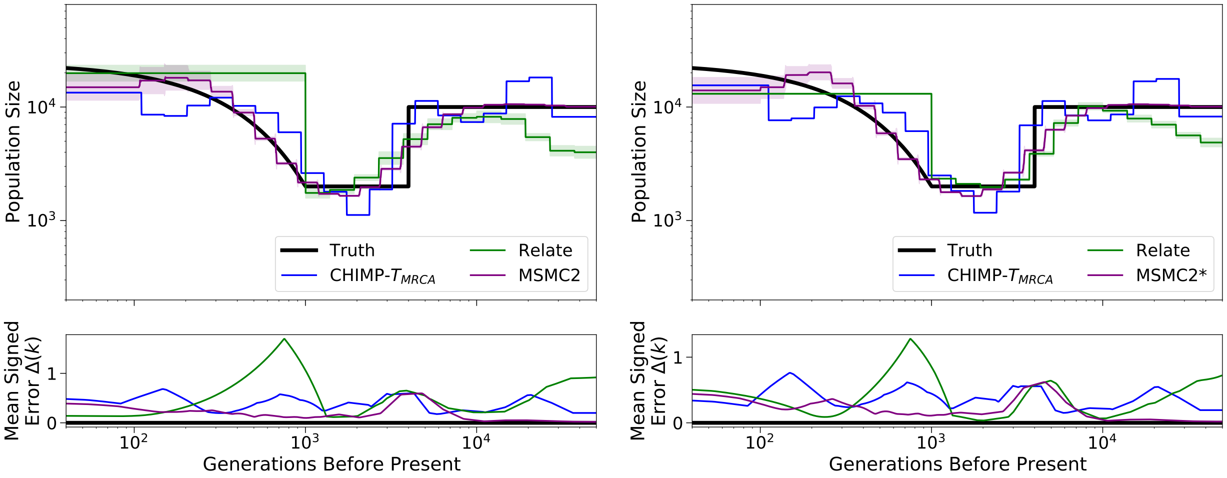
We also found some dependence on the size of the simplex. Recall that the simplices are defined in a space corresponding to $\text{Log}(\lambda_i)$, where λ_i are the coalescence rate parameters for each epoch i. A step size of $(1, 0, 0, 0...)$ then corresponds to multiplying the most recent coalescent rate λ_0 by e and leaving the other epochs the same. Since we would like our method to search widely early on in the inference procedure and narrow in later, we initializing the simplices in a shrinking fashion. Thus for the first EM iteration, the equilateral simplex is initialized to have larger edge lengths. For each successive iteration it is initialized to have progressively smaller edge lengths. Note that this procedure only describes the initialization of the simplex, and that within each M-step the simplex is further transformed as part of the Nelder-Mead maximization.

3.4.4 Heuristics for Discretization of Population Size History

Figure 3.7 shows the results when inferring the population size history in the bottleneck followed by exponential growth scenario (see Section 3.2.3), when using the default discretization for the different methods. For the method `CHIMP`, the default implementation chooses the discretization as follows: first we compute $\hat{N}$ from Watterson's estimator, and then partition the interval $[\frac{\hat{N}}{50}, 5 \times \hat{N}]$ (in generations) into 18 logarithmically equidistant epochs with an additional epoch being added above and below the minimum and maximum. Choosing an inappropriate discretization can have an impact on the accuracy of the inference, and can also lead to the inference missing important features of the underlying history.

3.4.5 Simulations with Recombination Map

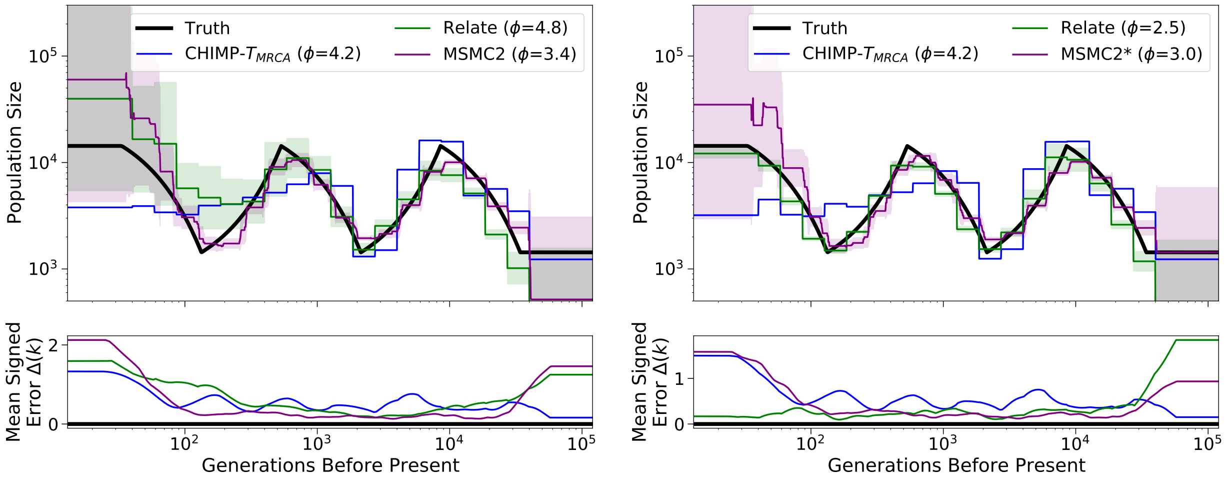
Here, we explore the effects of varying recombination rates along the chromosome on demographic inference using the different methods. We used `msprime` to simulate data under the *sawtooth* demography (explored in Section 3.2.3) using the HapMap II recombination rates on Chromosome 3 [39]. We inferred the population size history using `CHIMP`-T_{MRCA}, `MSMC2`, and `Relate`, and compared their performance using a the same discretization that was used



(a) Sample size 10.

(b) Sample size 200.

Figure 3.7: Results of inference in the bottleneck followed by growth scenario for sample size (a) 10 and (b) 200. We compare the results of CHIMP-T_{MRCA}, MSMC2, and Relate using the respective default discretization of the population size history. Truth shown in black. Solid lines are averages over 16 replicates and shaded area indicates standard deviation. Mean signed error $\Delta(k)$ is shown in bottom plot and has been smoothed using moving average for visualization purposes and the integral ϕ is indicated in the legend. (*) For sample size 200, MSMC2 was run on 50 non-overlapping pairs.



(a) Sample size 10.

(b) Sample size 200.

Figure 3.8: Results of inference for the sawtooth demography simulated using a human recombination map (for Chromosome 3) for sample size (a) 10 and (b) 200. We compare the results of CHIMP-T_{MRCA}, MSMC2, and Relate. For Relate we provide the recombination map used to simulate the data, for CHIMP we specify the constant recombination as the mean over the recombination map, and MSMC2 is set to infer the best fitting constant rate. Truth shown in black. Solid lines are averages over 16 replicates and shaded areas indicate standard deviation. Mean signed error $\Delta(k)$ is shown at bottom plot and has been smoothed using moving average for visualization purposes. The integral ϕ is indicated in the legend. (*) For sample size 200, MSMC2 was run on 50 non-overlapping pairs.

in Figure 3.4. We provided the methods with as much information about the recombination rates as possible: we provided Relate with the recombination map used to simulate the data. For CHIMP-T_{MRCA}, we specified a constant recombination rate corresponding to the mean recombination rate. Lastly, MSMC2 was set to infer a recombination rate while optimizing the population size history.

The results of these inferences are shown in Figure 3.8. Comparing these results to those in Figure 3.4, we note that CHIMP-T_{MRCA} and MSMC2, while operating under the assumption of a constant recombination rate, perform similarly well as in Figure 3.4, where the recombination rate was in fact constant. We also note that Relate, especially for 200 haplotypes, performs much better in the case of varying recombination rates. The reason for this improvement is likely the fact that the procedure used in Relate to infer genealogical trees can be shown to infer the correct tree in regions without recombination. Since there are many *cold* spots in the human recombination map, `Relate` infers the true underlying genealogies more accurately in these regions, and this yields an overall improvement in performance.

CHAPTER 4
DEMOGRAPHIC INFERENCE UNDER MODELS WITH POPULATION STRUCTURE

4.1 Introduction

The method presented in Chapter 2 and the results in Chapter 3 are for a single population model, where the population in question is assumed to be panmictic (freely mixing, where all individuals are equally likely to be the parents of the succeeding generation). However, many real world examples have the added complexity of population structure, where the assumption of a panmictic population begins to break down. These scenarios can often be modeled by demographies where there are multiple populations that coexist in time (and even space), that are each internally panmictic. Such demographies can include admixture events, where individuals are transferred between populations, or populations can split and merge. They also often include migration rates that allow for continuous mixture between distinct populations, helping emulate realistic behavior.

In this chapter, we introduce `CHIMP-S`, which reappropriates the ODE machinery used for computing $\mathbf{A}$ and $\mathbf{B}$ and uses it to solve for the probabilities in a population-split scenario, where an ancestral population splits at some point into two extant populations. We also demonstrate the inference of basic demographic parameters under using `CHIMP-S`.

4.2 Adapting `CHIMP` for a Population-Split Model

In Figure 4.1A, we see a diagram for the population-split model which we tackle with `CHIMP-S`. We infer demographic parameters for the size history functions $N_1(t)$, $N_2(t)$ and $N_A(t)$ in addition to the divergence time T all simultaneously, noting that $N_1(t)$ and $N_2(t)$ span times from $[0, T)$, while $N_A(t)$ spans $[T, \infty)$. The emissions off our model are the ordered pairs of derived alleles across the samples at each site, so we denote the vector of observations across the genome as $\vec{d} = \big((d_1^1, d_2^1), (d_1^2, d_2^2)...(d_1^L, d_2^L)\big)$, where the subscripts refers to the population in whose samples the derived alleles are observed, and the superscript indexes the position along the genome. The states s_i are defined as they were in the single population model, indicating which bin the T_{MRCA} for the entire two population genealogy falls into (we only focus `CHIMP-S` on T_{MRCA}, and ignore $\mathcal{L}$). Here we show how we compute the transition matrix **A** and emission matrix **B** for `CHIMP-S`.

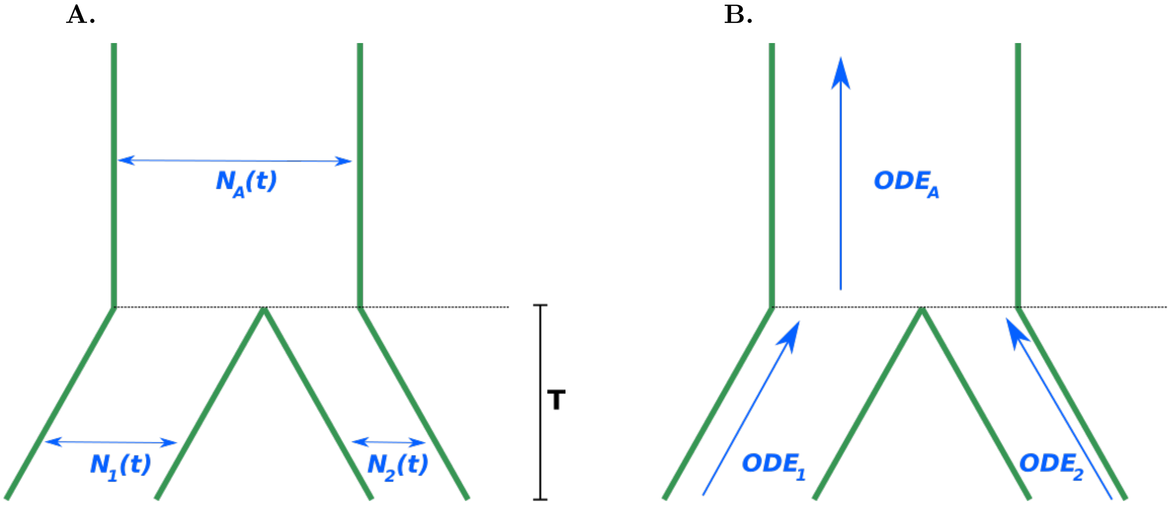


Figure 4.1: Schematic for (a) the demographic model under which CHIMP-S performs inference and (b) the ODE scheme which we employ under this model. The (a)panel displays the ancestral population which splits into two populations at some time T in the past. Each of the branch populations remain till the present day, where the samples are taken. The (b) panel shows how independent ODEs are run backwards in time to compute the evolution of each population independently. At time T, the states that ODE_1 and ODE_2 can take on are permuted and mapped to appropriate states of ODE_A, which evolves back in time until the absorbing state (when all samples reach the MRCA) is reached.

4.2.1 Computing the Transition Matrix **A**

The transition matrix **A** and emission matrix **B** are computed using a scaffolding of ODEs to compute the dynamics of the ancestral process in each sub population, and in the ancestral population separately (depicted in Figure 4.1B). Here we describe how this is done for the elements of the transition matrix $\mathbf{A}$. The ancestral process with recombination $\mathcal{A}^\rho$ that we use here is the same as the one used in the single population case. This means that the ODE and the numerical solution scheme is the same as before as well. In total, for each of the extant populations and the ancestral population, we have three such ODEs

$$\frac{d}{dt}\vec{g}_i^\rho(t) = \vec{g}_i^\rho(t) \cdot \mathbf{Q}_i^\rho(t), \tag{4.1}$$

where $i \in \{1, 2, A\}$. Note that while the structure of the ODEs is the same for each sub-population, the evolution matrix $\mathbf{Q}_i^\rho(t)$ will differ between the populations because the values and structure of Q depends on the total number of samples in the ancestral process for a particular sub-population and its population size history for that population. We begin solving the transition matrix by running ODE_1 and ODE_2 back in time until T. In other words, we integrate equation (4.1) from $\int_0^T dt$ to solve for $\vec{g}_1^\rho(T)$ and $\vec{g}_2^\rho(T)$. Note that $\mathbf{Q}_1^\rho(t)$ and $\mathbf{Q}_2^\rho(t)$ are defined with $N_1(t), n_1$ and $N_2(t), n_2$, respectively (n_1 and n_2 are the numbers of haplotype samples collected at present day for populations 1 and 2). The initial conditions for both of these processes are the states corresponding to n_1 and n_2 lineages ancestral to both loci A and B, that is, in the notation of the previous chapter, $\vec{g}_1^\rho(0) = (n_1, 0, 0, 0)$ and $\vec{g}_2^\rho(0) = (n_2, 0, 0, 0)$

Next, the vectors $\vec{g}_1^\rho(T)$ and $\vec{g}_2^\rho(T)$ are used to compute the initial values for ODE_A (i.e. $\vec{g}_A^\rho(T)$). This is done by mapping ordered pairs of states $\big(\sigma_1, \sigma_2\big)$ to states (σ_A). This mapping represents the joining of the two populations, and recalling that the states of the ODE's σ are indexed by tuples of 4 numbers (k_{ab}, k_a, k_b, r), in mapping (σ_1, σ_2) to σ_A, elements of the tuples are all summed independently, representing summing the lineages of each type across the two populations at time T. Since this introduces certain states in the ancestral population which may have had two recombinations, (from summing $r_1 = r_2 = 1$, we set the probability density in these states to zero ($g_A^{(*,*,*,2),\rho}(T) = 0$), and renormalize over the remainder of states such that the sum of probability equals 1. Ignoring states with two recombinations and renormalizing is equivalent to assuming that the recombination rate is small, and helps computational tractability. Thus, we define the mapping

$$\big((k_{ab,1}, k_{a,1}, k_{b,1}, r_1)_1, (k_{ab,2}, k_{a,2}, k_{b,2}, r_2)_2\big) \rightarrow \begin{cases} (k_{ab,1} + k_{ab,2}, k_{a,1} + k_{a,2}, k_{b,1} + k_{b,2}, r_1 + r_2)_A, & \text{if } r_1 + r_2 \leq 1 \\ \text{None}, & \text{otherwise} \end{cases} .$$

Note that ODE_A is defined to allow for a total of $n_A = n_1 + n_2$ lineages, allowing it to capture as many of the ordered pairs of the merging ODEs as necessary. With this mapping, we can write down the full initial condition for ODE_A as

$$g_A^{\sigma_A,\rho}(T) = \frac{1}{z} \sum_{(\sigma_1,\sigma_2)\rightarrow\sigma_A} g_1^{\sigma_1,\rho}(T) g_2^{\sigma_2,\rho}(T),$$

where the sum is over ordered pairs that map to the respective state σ_A, and the renormalization constant z is equal to the sum of the numerator over all viable states σ'_A.

Thus we arrive at the initial condition for ODE_A, and integrate equation (4.1) again from $\int_T^t dt$, to compute $\vec{g}_A^{\rho}(t)$. Then, as in the single population case, we use equation (2.5) to obtain the joint cumulative transition probabilities, from which the transition matrix is computed as before. Note that because the most recent possible T_{MRCA} is the divergence time T, the probabilities of transitioning into, out of, and between states below T is 0.

4.2.2 Computing the Emission Matrix **B**

The method for computing the emission probability matrix **B** for the single population case is adapted to the population split model in much the same way as was demonstrated for the transition probabilities. The ancestral process with mutation $\mathcal{A}^{\theta}$ is run separately for the populations $i \in \{1, 2, A\}$, using three separate ODEs. At the divergence time T, the ordered pairs of states (σ_1, σ_2) belonging to ODE_1 and ODE_2 are mapped to appropriate states of ODE_A, (cases where mutation has already occurred twice are ignored, and the probabilities are renormalized across all valid states). In computing the cumulative emission probabilities, a number of additional combinatorial factors are used to account for various configurations of derived alleles that subtend a mutation mapping onto the appropriate emission states (d_1, d_2).

4.3 Demographic Inference with `CHIMP-S`

We used the transition and emission probability matrices, as computed above, in an EM inference framework, similarly to the single population case in Chapter 3, to infer demographic parameters. Forty replicates of data were simulated using `msprime`, for a population model where each of the extant populations, as well as the ancestral population, had constant sizes of 10000 individuals, and the populations diverged 4000 generations in the past. The per base, per generation recombination and mutation rates were set to $r = \mu = 10^{-7}$, and for each replicate 5Mbp of genome was simulated.

Then, using `CHIMP-S`, we inferred the constant ancestral size N_A and the divergence time T jointly, specifying the same mutation rate, recombination rate, N_1, and N_2 values that were used in the simulation. Figure 4.2 shows that we obtain reasonably accurate results for this scenario.

4.4 Conclusion

In this chapter we demonstrate the use of the underlying model of `CHIMP` in extensions that perform inference under models with population structure. We presented `CHIMP-S` a tool

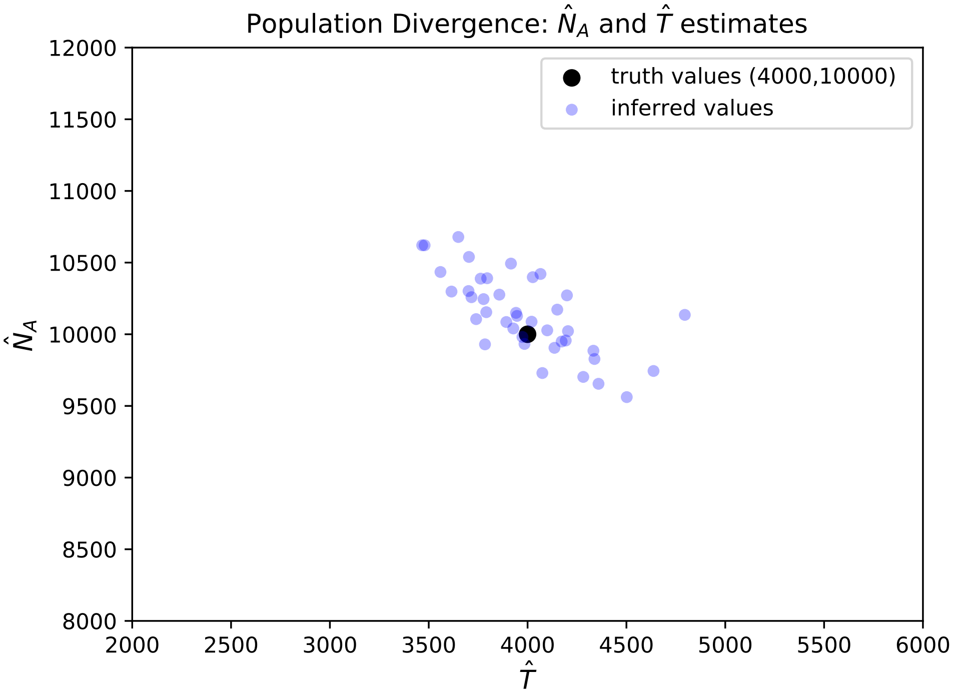


Figure 4.2: Results for two parameter inference using the split-population model of `CHIMP-S`. Estimates for 40 replicates are shown.

that analyzes data under a population-split model, and demonstrated how to compute the appropriate probabilities and perform inference with it. The simulated scenario and the corresponding inference was performed in a constrained setting where favorable results were achieved, and the next step to developing `CHIMP-S` into a full-fledged inference tool would be to conduct more extensive studies testing its robustness in different scenarios and comparing its results to those of other methods. Another highlight of `CHIMP`'s approach is the fact that the underlying ODE computational system can be adapted to handle different scenarios. For example, in much the same way that we handle split-populations with `CHIMP-S`, we foresee adapting the ODE's to handle pulse-admixture events. We could even incorporate asynchronous samples to allow for inference using ancient DNA samples in tandem with modern samples, possibly even with more than two populations. Another direction `CHIMP` can be developed is to handle cases of continuous migration; in this case, there are complexities arising from the populations interacting continuously that would require additional model extensions. Even if more simplifying assumptions are made to the computational scheme that make the method less exact, the fact that `CHIMP` is compatible with unphased and pseudohaploid data could prove for this to be a fruitful direction for investigation.

CHAPTER 5

DECODING THE POSTERIOR DISTRIBUTION ACROSS STATES

5.1 Motivation

While much of the CHMM framework so far has assumed that the genome evolves neutrally, in reality, selection plays a role in evolution, leaving patterns of adaptive genetic variation. This form of variation is often the central subject of medical genetic studies that seek to establish causal links between regions of the genome and phenotypic traits, and thus, methods that offer new ways to detect and understand patterns of adaptive variation are an important part of the field.

While CHMM methods have primarily been used to infer the population size history, here we explore further downstream studies that can be conducted using the posterior distribution of the hidden states, which are a byproduct computed during the inference process. The posterior probabilities describe the state the model is in sequentially along the genome, and can provide scientists with additional information about the evolutionary processes that are shaping the genetic information, especially if the posterior is markedly different from what would be expected under the null model of a single neutrally evolving population. In particular, adaptive genetic variation, or selection, can leave signatures where the patterns of variation significantly differ from the neutrality. To this end, we developed `CHIMP-PD`, a tool that uses the CHMM framework to produce an estimate of the posterior probabilities that can be interpreted and used in downstream studies of adaptive variation.

We focus on two archetypes of adaptive variation that are brought about by i) selective sweeps, and ii) balancing selection. `CHIMP-PD` is well-suited to uncover evidence of these modes of selection because of the effect they have on the shape of ancestral trees. In a selective sweep arising from a beneficial allele, all alleles in the genetic sample can be traced to a beneficial allele that has quickly established and fixed itself, therefore in such a region, the trees are often much shorter than under neutrality. Conversely, for an region that displays balancing selection, multiple variants are maintained in the population for long periods of time at intermediate frequencies. As a result, the common ancestor of the two allele types is found in the deep past, and the underlying trees are very tall.

In this chapter we begin by showing how we obtain the posterior probabilities using `CHIMP-PD` and evaluate the accuracy of these estimates in neutral simulations. Then we study the aforementioned archetypes of adaptive variation, selective sweeps and balancing selection, and demonstrate their detection using `CHIMP-PD`. We focus on simulated genetic data first, and then examine examples in real human data.

5.2 Computing Posterior Probabilities

For `CHIMP-PD`, the posterior distribution across states is computed during the forward backward algorithm as described in Section 2.5.1, and is given by the quantity $\gamma_i(\ell)$ defined in Equation (2.19). This tells us the probability that the hidden state of the CHMM is in each state s_i and is computed for each locus (or meta-locus) along the genome. From these probabilities we can investigate the T_{MRCA} and $\mathcal{L}$ across the genome since the hidden states model the behavior of these variables.

For large samples, this genomic sequence of posterior probabilities is computed independently for each subgroup of n_s samples in accordance with our composite likelihood framework. While the composite likelihood is straightforwardly computed as a sum of the log-likelihoods of the data for each subgroup of samples, consolidating the information from the posterior probabilities for each subgroup requires some additional steps. We opt to average the posterior probabilities across the subgroups, yielding an estimate for the posterior probability of a random sample of n_s haplotypes drawn from the population

$$\bar{\gamma}_i(\ell) = \frac{1}{\lfloor \frac{n}{n_s} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{n_s} \rfloor} \mathbb{P}[s^\ell = s_i | \vec{d^i}].$$

It is important not to confuse this with an estimate for the actual posterior distribution of the state variable across all the samples, or even the posterior for a particular group of samples. However, reconstructing an accurate estimate of T_{MRCA} or $\mathcal{L}$ can be achieved choosing $n_s = n$ if desired.

5.3 Simulations Under Neutrality

While Chapter 3 shows the efficacy of `CHIMP`'s model in performing demographic inference, we would like to more specifically evaluate the ability of the model to accurately infer the T_{MRCA} and $\mathcal{L}$. To this end, we conducted a series of simulations where the true values of the T_{MRCA} and $\mathcal{L}$ are retained from the simulation, and compared to values inferred using both `CHIMP-PD` and `Relate` [10, v1.1.3].

The simulations were conducted using `msprime` [25, v0.7.4]. We simulated 16 replicates of data, where each replicate contains 200Mbp of genomic data for 10 haplotypes, simulated for a constant population size $N = 10000$ with $r = \mu = 1.25 \cdot 10^{-8}$ to mimic human parameters.

When analyzing these data with `Relate`, we specified the true constant population size and infer the tree sequences. Then we computed T_{MRCA} and $\mathcal{L}$ for these estimated trees based on `Relate`'s inferred tree output.

For the `CHIMP-PD` analysis, the posterior probabilities were computed using a constant population size history based on Watterson's estimator (as in Section 3.4.1). There is some ambiguity in estimating the state variable T at locus ℓ (representing T_{MRCA} or $\mathcal{L}$) given $\bar{\gamma}_i(\ell)$, since each state i indicates a range $t_{i-1} \leq T_\ell < t_i$ for the state variable T (as in section 2.2).

To this end, we associate a single intermediate value $\hat{t_i}$with each state i, computed as the arithmetic mean of the bounds of the state (except for the last state S which is unbounded and requires an adjustment) given by

$$\hat{t_i} = \begin{cases} 2t_{S-1} - t_{S-2}, & \text{if } i = S \\ \frac{t_{i-1}+t_i}{2}, & \text{otherwise.} \end{cases} \tag{5.1}$$

In the following analysis, we compare the results of using three straightforward estimates based on the mode, median, and mean of the distribution, computed as follows. Based on $\hat{t}$'s defined in equation (5.1), and given the posterior $\bar{\gamma}_i$ for a given locus ℓ, we can define the mode, median, and mean of T as

$$\begin{aligned} T_{mode} &= \hat{t_j}, \text{ where } j = \underset{i}{\operatorname{argmax}} \left[\bar{\gamma}_i \right] \\ T_{median} &= \hat{t_j}, \text{ where } j = \inf \left\{ k \middle| \sum_{i=0}^{k} \bar{\gamma}_i \geq 0.5 \right\} \\ T_{mean} &= \sum_i \bar{\gamma}_i \cdot \hat{t_i}. \end{aligned}$$

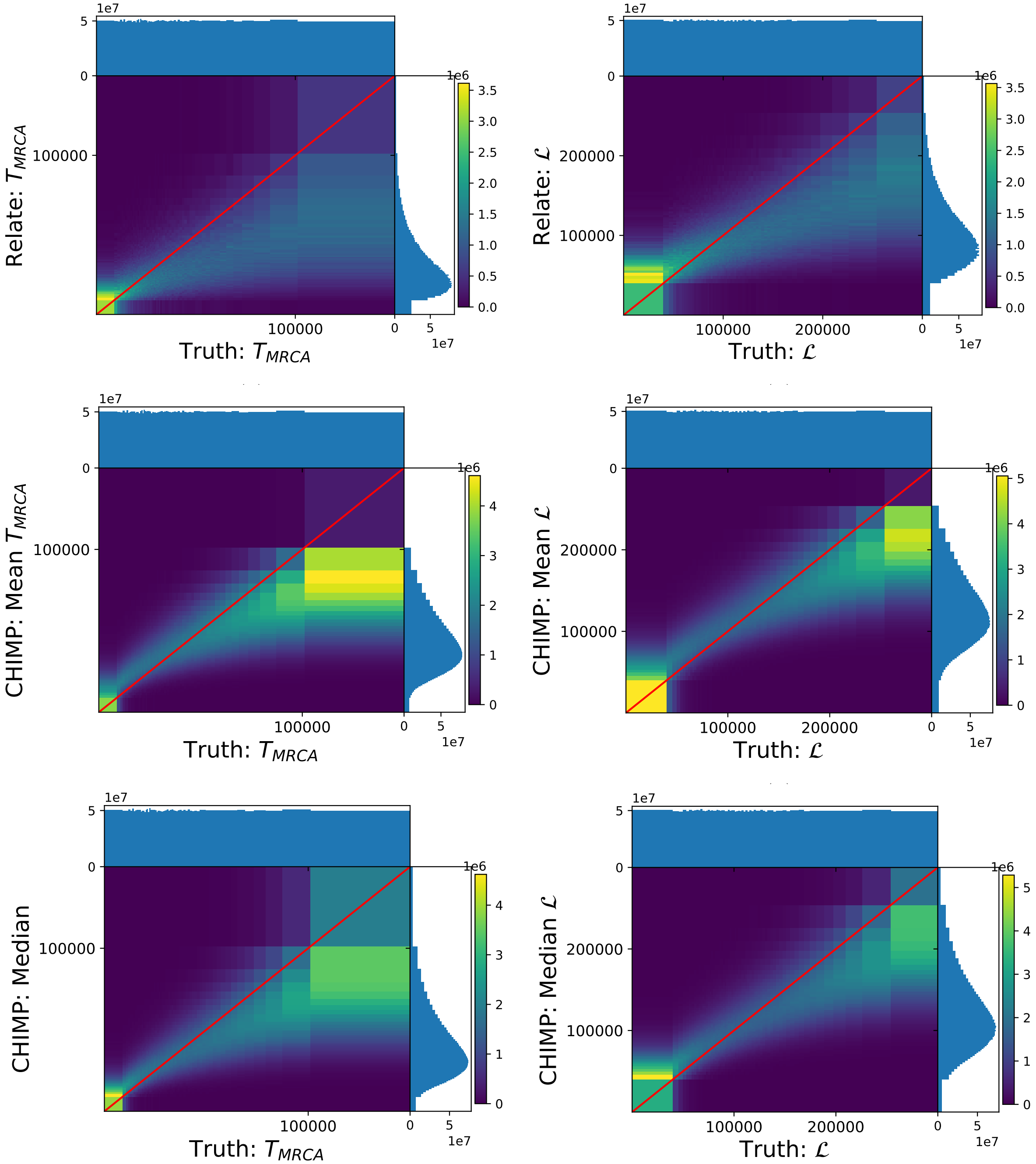
With these estimates for T_{MRCA} and $\mathcal{L}$, we compare the profile of estimated values to the profile of true values and present the results in figures 5.1 and 5.2. Each individual plot contains aggregate data from across all 16 replicates of the simulations, ie. a total of 3.2 Gbp of data for 10 haplotypes. The heat map compares the estimate from `Relate` or `CHIMP-PD` to the true value of T, giving equal weight to each base on the genome (not each inferred tree). In addition to a heat map, we also include one dimensional histograms of the estimates along the borders.

Figure 5.1 contains the comparison of the truth to results from `Relate` and `CHIMP-PD`'s mean and median estimates. Both the heatmap and the adjacent 1-D histograms are binned according to the bins that CHIMP-PD uses for internal states. This can be seen in the roughly constant distributions of true T_{MRCA} and $\mathcal{L}$ since CHIMP-PD chooses its states such that the marginal distribution across them (based on a prior of constant population size) is constant.

The 1-D histograms show that CHIMP-PD is more accurate in recovering the true (constant) distribution. The distributions for Relate are more highly peaked, and do not give enough weights to the tails of the distribution. This alone does not show that CHIMP-PD is performing better, since this distribution could in principal be achieved with results that are uncorrelated to the true values. However, the heat maps allow us to visualize this correlation. The red line indicates where the estimated T values match the truth, and perfect recovery of the true T values would be reflected by a heat density that is uniformly spread along this line. By this measure, CHIMP-PD is more accurate at inferring T than Relate, since the density for CHIMP-PD is more concentrated along the red line. The high density ridge for Relate has a slope less than 1, which indicates that Relate is conservatively inferring values of T that are closer to a genome-wide average, whereas CHIMP-PD more accurately estimates T, especially for very high or low values.

CHIMP-PD's mean and median estimates perform similarly to each other, though the median estimates seem better able to infer very deep trees. This makes sense given the way we compute median and mean values from the posterior probabilities, since the fact that the mean is a weighted sum means that probability density in the lower states can pull the estimate lower whereas the median can more readily return $\hat{t_S}$ as an estimate. Using $\mathcal{L}$ as the state variable yields distributions that match the truth better than T_{MRCA}, and this is evidenced in the right column where using $\mathcal{L}$ produces heat maps that are closer to the red line, and 1-D histograms with wider distributions.

In Figure 5.2 we present the analogous comparisons using T_{mode} estimates from CHIMP-PD. To better visualize the results, each bin for the heatmap and histograms is formed by grouping



(e)

(f)

Figure 5.1: Correlations between true and inferred values for the state variables T_{MRCA} (left) and $\mathcal{L}$ (right), comparing inference using Relate and CHIMP (separately using the mean and median of CHIMP's posterior distribution as the estimator). The results are aggregated across 3.2 Gbp of simulated data (10 haplotypes simulated under constant 10k population size with human r and μ using msprime). We also include marginal 1-D histograms for inferred and true values. Both heatmaps and histograms are binned according to CHIMP's hidden states.

4 hidden states of CHIMP-PD, which still preserves the uniform distribution across bins of the true T values. A log scale is used to more clearly see the crest of high density for the heat maps. We see that using CHIMP-PD's T_{mode} estimates results in a bimodal distribution that is more heavily weighted at the tails, nominally the opposite problem that Relate has. Indeed, we can see that the slope of the high-density crest of the heatmap is less than one in the Relate plots, and is greater than one in the T_{mode} plots, underscoring the tendency of Relate to be biased downwards from the mean and of CHIMP-PD T_{mode} to be biased upwards.

Figures 5.1 and 5.2 show that, overall, CHIMP-PD has more flexibility than Relate to accurately recover T_{MRCA} and $\mathcal{L}$ along the genome. To quantify this performance, we compute the deviation of the estimates from the true values of T_{MRCA} and $\mathcal{L}$. We define

$$\sigma = \sqrt{\frac{i=1}{L}\sum_{i}^{L}(\log \hat{T}_i - \log \bar{T}_i)^2}$$

where i indexes the sites simulated ($L = 3.2$ Gbp in total), and $\hat{T}_i$ and $\bar{T}_i$ denote the estimated and true values, respectively, of the state variable at a given site i. This measure of average error across the simulated data for each of the methods is collected in table 5.1. The σ values show that CHIMP-PD, particularly with the mean and median estimates, is much more accurate than *Relate* at recovering the true values of T_{MRCA} and $\mathcal{L}$.

In Figure 5.3 we show the actual estimates for T_{MRCA} and $\mathcal{L}$ for a stretch of simulated genome. We omit the median estimates from CHIMP-PD due to their similarity to the mean estimates. We see that when the true values of T dips very low, the CHIMP-PD and Relate do similarly well at recovering T, however when T takes on high values indicating very deep trees, inference with CHIMP-PD is much more accurate. We also see the different behaviors of CHIMP-PD's T_{mode} and T_{mean} estimates. The mean estimates take a more "nuanced"

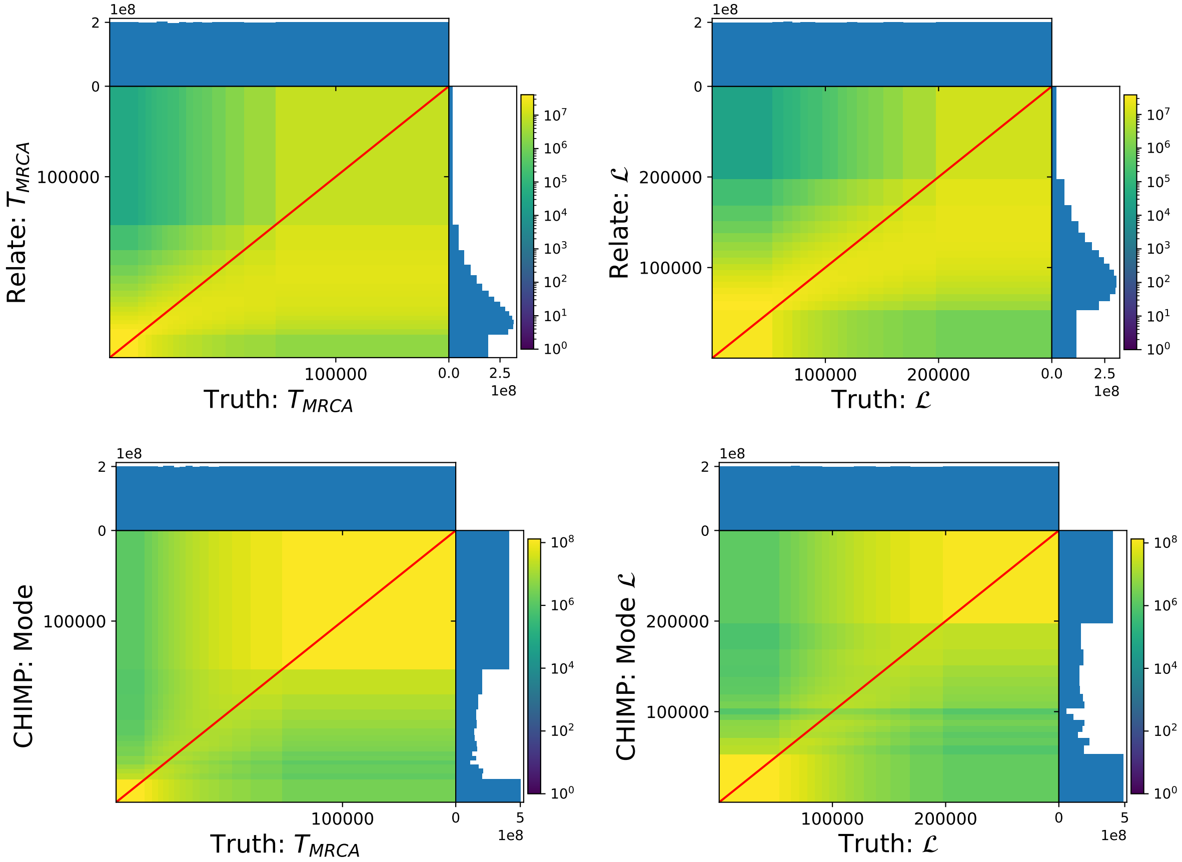


Figure 5.2: Correlations between true and inferred values for the state variables T_{MRCA} (left) and $\mathcal{L}$ (right), comparing inference using `Relate` and `CHIMP`'s mode estimator. The results are aggregate across 3.2 Gbp of simulated data (10 haplotypes simulated under constant 10k population size with human r and μ using `msprime`). As before, we include marginal 1-D histograms. To help visualize the density crest, we use a coarse binning in the heatmap such that every 4 consecutive hidden states are considered as 1 bin. These bins are also used for the histograms.

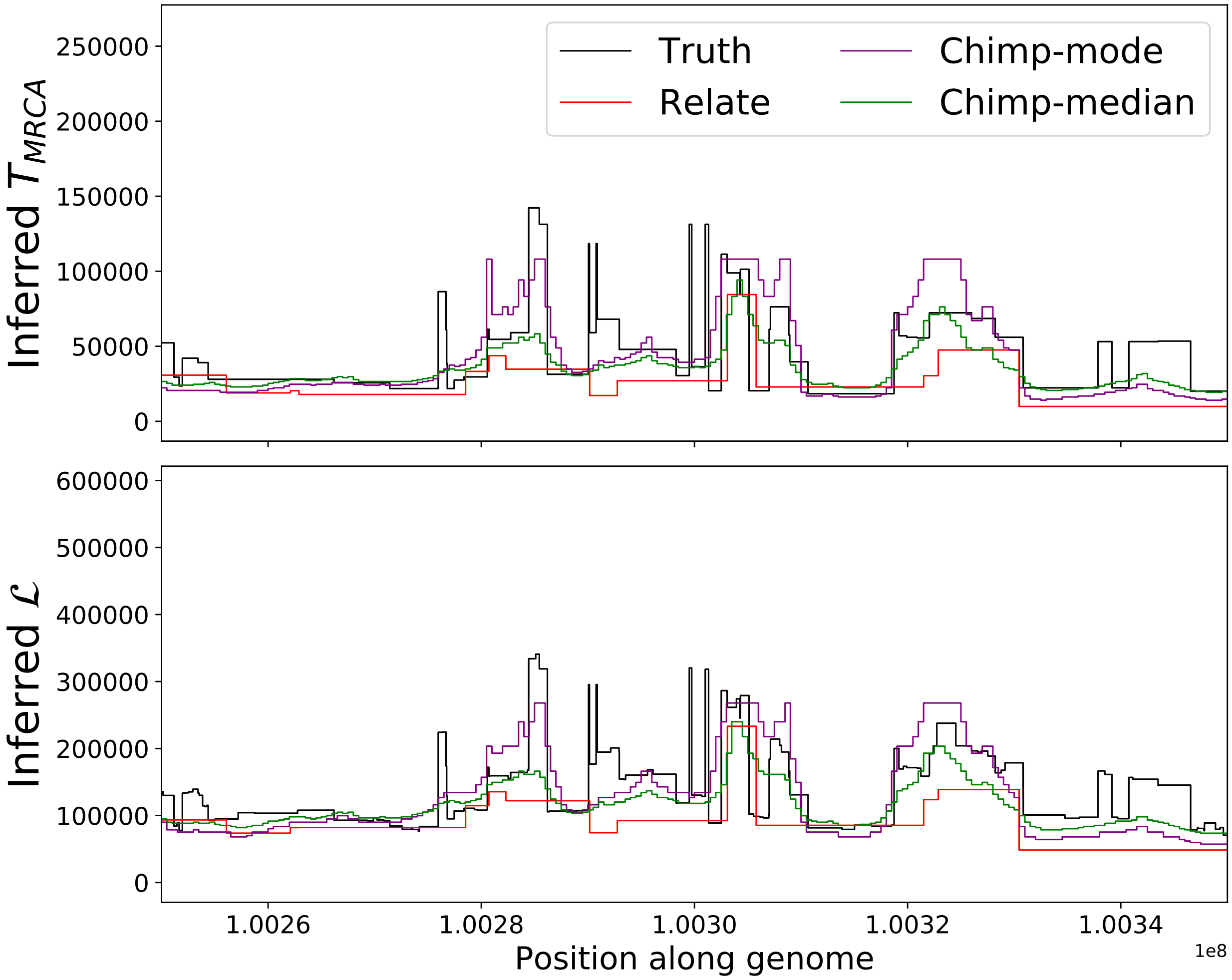


Figure 5.3: Example of inferred T_{MRCA} (top) and $\mathcal{L}$(bottom) along a 100 kbp stretch of simulated data under neutrality, with the true values included for comparison. The data was a segment sampled from the 3.2 Gbp used in Figures 5.1 and 5.2. For `CHIMP` we omit the median estimator for visual clarity as it is very similar to the mean. The apparent upper bound on the values inferred by `CHIMP` is an artifact of the fact that the HMM uses discrete states and we have assigned a finite value to the last state.

Table 5.1: Comparison of accuracy (σ) for different estimates of the state variable. Recall $\sigma = 0$ corresponds with perfect accuracy.

Estimated Quantity	Method	σ for 3.2 Gbp data
T_{MRCA}	Relate	0.584
	CHIMP:Mean	0.431
	CHIMP:Mode	0.648
	CHIMP:Median	0.425
$\mathcal{L}$	Relate	0.409
	CHIMP:Mean	0.317
	CHIMP:Mode	0.485
	CHIMP:Median	0.315

approach, however the mode estimates often overestimate or underestimate the true features. Figure 5.3 visually reinforces the findings from table 5.1, showing that CHIMP-PD's mean estimator is much more accurate than Relate, and while CHIMP-PD's mode estimator is less accurate overall, it does capture the tails (high and low values) of the distribution of the state variable better.

5.4 Simulations Under Selection

Now that we have shown that CHIMP-PD accurately recovers the T_{MRCA} and $\mathcal{L}$ in neutral simulations, we shift our focus to detecting adaptive genetic variation. In this section we simulate and analyze scenarios of selection using SLiM [40], a forward-time simulation tool that is well suited for selection. The two types of selection we examine are selective sweeps, where a beneficial allele rises in frequency and fixes, and balancing selection, where heterozygote-advantage maintains roughly equal frequencies for two different alleles at a given locus.

5.4.1 Selective Sweep

For our simulations of the selective sweep scenario, we simulated (using `SLiM`) the dynamics for a (constant-size) population of 5000 diploid individuals with $r = \mu = 1.25 \cdot 10^{-8}$ to mimic human-scale parameters and focused on a 3Mbp genetic region. During the simulation this population is allowed to "burn-in" for $5 \cdot 10^4$ generations, during which random neutral mutations accumulate along the genome and create a background of neutral genetic variation that resemble the background characteristic of mutation-drift equilibrium. After this burn-in period, a beneficial mutation $m*$ is introduced at a single base positioned in the middle of the 3Mbp window in one haplotype of the population. This mutation has a dominance coefficient $h = 1/2$ and a selection coefficient s that determines the strength of the selective advantage. The population is allowed to evolve after the introduction of the beneficial variant. If the mutation is lost (frequency falls to 0), the simulation is restarted after the burn-in period and $m*$ is reintroduced. This is repeated until the frequency of $m*$ sweeps to 1. At the time $m*$ is fixed, the full genetic data for 150 individuals are randomly sampled from the population. We conducted 16 replicates of this simulation for each value of the selection coefficient $s = 0.01, 0.02$, and 0.05 (recent studies have shown that $s \approx 0.02$ for lactase persistence in certain human populations [41]).

After simulating this data, we ran `CHIMP-PD` on each replicate to obtain the posterior distribution. We computed the mean, mode, and median for this distribution at each metalocus as described in Section 5.3. For this analysis, `CHIMP-PD` assumed a constant population-size history based on Watterson's estimate computed from the data, and used a metalocus size of 500 bases. We anticipate that if the beneficial allele was introduced fairly recently and then sweeps to fixation, the T_{MRCA} and $\mathcal{L}$ near the selected locus will reflect very short trees. So rather than using hidden states defined by an equal probability of occupancy under neutrality, we broke up the state corresponding to the shortest tree into smaller states at very recent times to allow `CHIMP-PD` to infer very short trees with more precision. Breaking

Table 5.2: Results from analyzing selective sweep simulations with `CHIMP-PD`.

Selection Strength	State Variable	True Signals Observed (out of 16)	False Signals Observed
$s = .01$	T_{MRCA}	10	0
	$\mathcal{L}$	10	0
$s = .02$	T_{MRCA}	15	1
	$\mathcal{L}$	15	1
$s = .05$	T_{MRCA}	16	1
	$\mathcal{L}$	16	1

up the smallest state in this way also means that the marginal probability of the CHMM occupying these states is decreased, so that actually inferring these states constitutes strong evidence of systematically small trees.

We analyzed all 16 replicates of data for each $s = 0.01, 0.02$, and 0.05 and tabulate the results in Table 5.2. For each s value, we record the number of selection signals we observe across the 16 replicates using both T_{MRCA} and $\mathcal{L}$. To classify whether an observed pattern of inferred T constitutes a "signal", we visually evaluate whether the mode of the posterior distribution is among the additional states that were added in the very recent past to increase precision ($< 6k$ generations for T_{MRCA} and $< 20k$ generations for $\mathcal{L}$), and mark any such contiguous region in the 3Mbp of simulated data as an observed "signal". Note that the regular states have a marginal occupancy probability of 0.02, but these additional states have probabilities < 0.02 under the prior; this makes them easy to visually identify at the bottom of the posterior density plots we show. We use the mode of the posterior for determining whether a signal is observed because of its heightened sensitivity to deviant behavior of the state variable (seen in Section 5.3), though in the plots we show the mode, median, and mean. In our results there does not appear to be a significant difference between T_{MRCA} and $\mathcal{L}$. We see that `CHIMP-PD` can easily uncover the signals of strong selection,

($s = .02$ and $s = .05$), however for weaker selection ($s = .01$), it becomes more difficult. Table 5.2 also tabulates the number of "false signals" we observed across all 16 replicates of data, ie across $48Mbp$ of simulated genome. These are patterns in the data that we classify as signals based on our criterion, but that do not occur in regions that overlapping with the selected allele. Though a few such false signals appear in these simulations, we note that they occur very close to the true signal and are likely due to ancestral recombination events that fall in such a way as to excise some portion of the signal.

We show a few illustrative examples of the posterior obtained with CHIMP, with the mean, median, and mode computed. Figure 5.4 shows the posteriors for a dataset simulated with $s = 0.2$. The mutation under selection is at position $1.5Mbp$, and we see high probabilities of finding very small T_{MRCA} and $\mathcal{L}$ values around this region. Near this position, each of the mean, median, and mode of the posterior are observed in the very recent states included in the recent past to provide more precision, taking on values much lower than the apparent floor that is observed in the remainder of the background distribution. Though in Table 5.2 there was no significant difference between the discriminatory power of T_{MRCA} and $\mathcal{L}$, here we see that the signal observed in the $\mathcal{L}$ distribution is wider, and thus more prominent, than that in the T_{MRCA} distribution. Though only a single base in the genome is under selection, the evidence of selection extends upstream and downstream of this region because of the genetic hitchhiking of other neutral mutations from the original haplotype on which the selected mutant originated. However, as more time elapses between the sweep and the sampling of data, further recombinations will shrink the region that remains correlated with the selected position, and the signal will grow weaker. The signal also appears weaker for weak selective sweeps, like that shown in Figure 5.5 where we do not observe any clear signal in the $\mathcal{L}$ distribution for a dataset simulated under $s = 0.01$. Our naive criterion using the mode of the posterior does not indicate any signal in Figure 5.5, though near the selected site there is visual evidence of $\mathcal{L}$'s that are on average lower than the background. A more

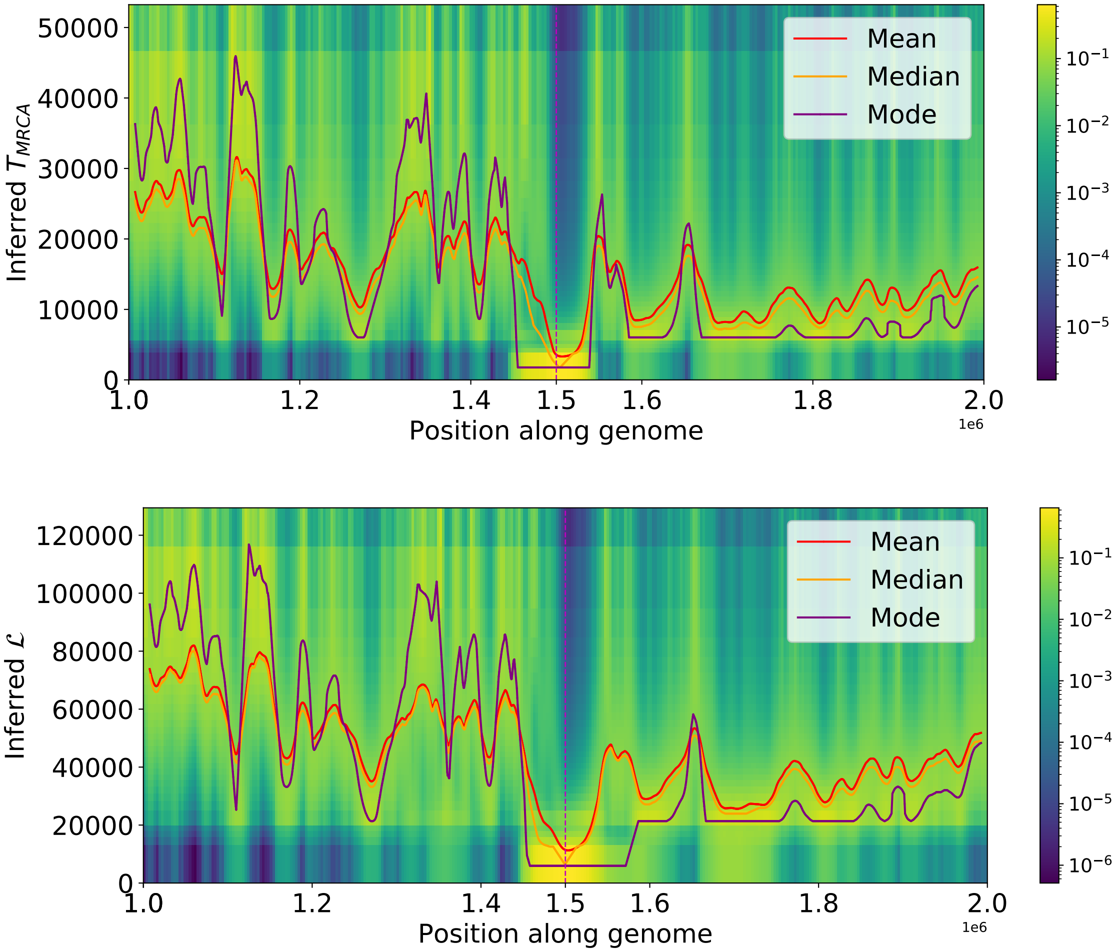


Figure 5.4: Posterior distributions computed by using CHIMP-PD to analyze data collected from a selective sweep simulation with $s = 0.02$. This is computed for both T_{MRCA} (top) and $\mathcal{L}$ (bottom). The genomic segment is centered around the beneficial mutation site. We further compute the mode, mean, and median of the distribution and overlay these curves. The regions below $T_{MRCA} \approx 6k$ and $\mathcal{L} \approx 20k$ generations are subtended by a finer division of states that were included to increase precision in the recent past, and we see that the mode, median, and mean all extend into the region near the selected mutation, showing clear evidence of a signal. The signal for $\mathcal{L}$ appears wider and more pronounced than that for T_{MRCA}. In the heatmap, brighter colors indicate higher inferred probability for the state spanning that region.

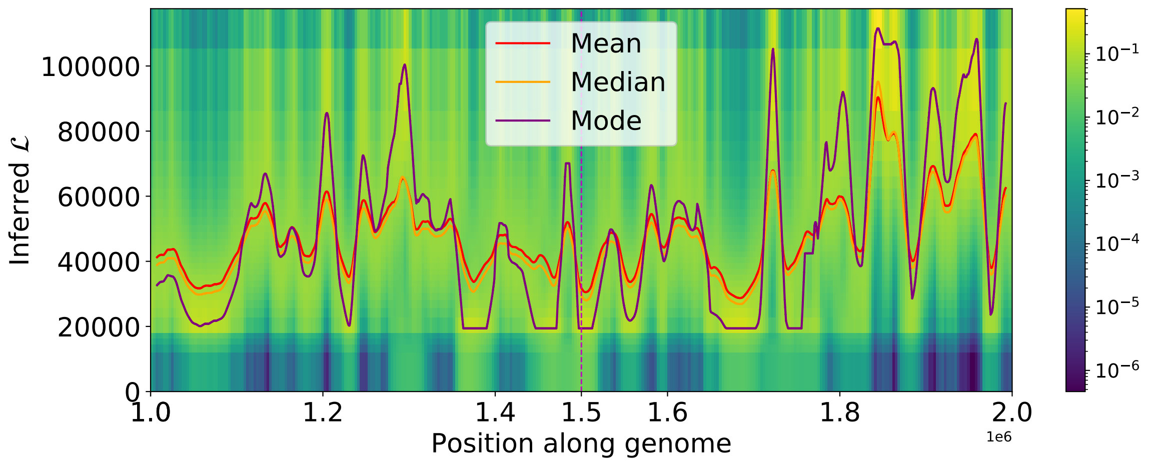


Figure 5.5: Posterior distribution of $\mathcal{L}$ as computed by using `CHIMP-PD` to analyze data collected from a selective sweep simulation with $s = 0.01$. We also display the mode, mean, and median of the distribution. By our criteria, we do not detect the selection since we do not see $\mathcal{L}$ extend into the recent states (below $\approx 20k$ generations), though we do observe a region with a generally low average value for $\mathcal{L}$.

sophisticated signal processing pipeline would likely enable us to more clearly detect this evidence of selection.

5.4.2 Balancing Selection

For our simulations of balancing selection, genetic variation in a region under selection is maintained by heterozygote-advantage of a haplotype. We use the same population size and demographic parameters as for the selective sweep, and the same burn-in period as well. We model selection on an extended region of the genome (as opposed to a single site) by introducing 21 evenly spaced mutations of type $m*$ on a single random haplotype from the population, and label this haplotype $h*$. The $m*$ mutations are spaced 1kbp apart from their neighbors, and in total span a 20kbp region that is centered in the middle of the 3Mbp window. Selection is defined to act such that an individual gains a fitness advantage of s only if they have one haplotype with 0 $m*$ mutations, and one with 21 $m*$ mutations. After introducing $h*$, the simulation proceeds for 10^5 generations before 150 individuals'

Table 5.3: Results from analyzing balancing selection simulations with `CHIMP-PD`.

Selection Strength	State Variable	True Signals Observed (out of 16)	False Positive Observed
$s = .01$	T_{MRCA}	16	0
	$\mathcal{L}$	16	0
$s = .02$	T_{MRCA}	16	0
	$\mathcal{L}$	16	0
$s = .05$	T_{MRCA}	16	0
	$\mathcal{L}$	16	0

data is collected. If at any point $h*$ is lost or fixes, the simulation is restarted after the burn-in. As before, we simulated a total of 16 replicates for each of the selection strengths $s = 0.01, 0.02,$ and 0.05, and analyzed these replicates using `CHIMP-PD`, this time breaking up the highest state (corresponding to the largest trees) into smaller bins to obtain better resolution in the distant past.

The results for all 16 replicates of data for each $s = 0.01, 0.02,$ and 0.05 are tabulated in Table 5.2. The criterion for determining whether a signal was observed for a given simulation replicate was too visually evaluate whether the mode of the posterior achieved high values in the additional states added in the distant past ($> 55k$ generations for T_{MRCA} and $> 135k$ generations for $\mathcal{L}$) for a consecutive stretch of data spanning $>$ 10kbp anywhere in the simulated genome, and to mark each such stretch as a "signal". As before, because the marginal occupancy probability distribution of these states under the prior is < 0.02, these states are visibly distinct in the density plot of Figure 5.6. For each of the selection coefficients we used, `CHIMP-PD` detects clear signals of balancing selection in the appropriate regions. While we do see some spikes in surrounding regions that somewhat resemble the profile of our primary signal (as in the $\mathcal{L}$ posterior in Figure 5.6), these spikes are likely caused by random SNPs in the background that happen to maintain intermediate frequencies for

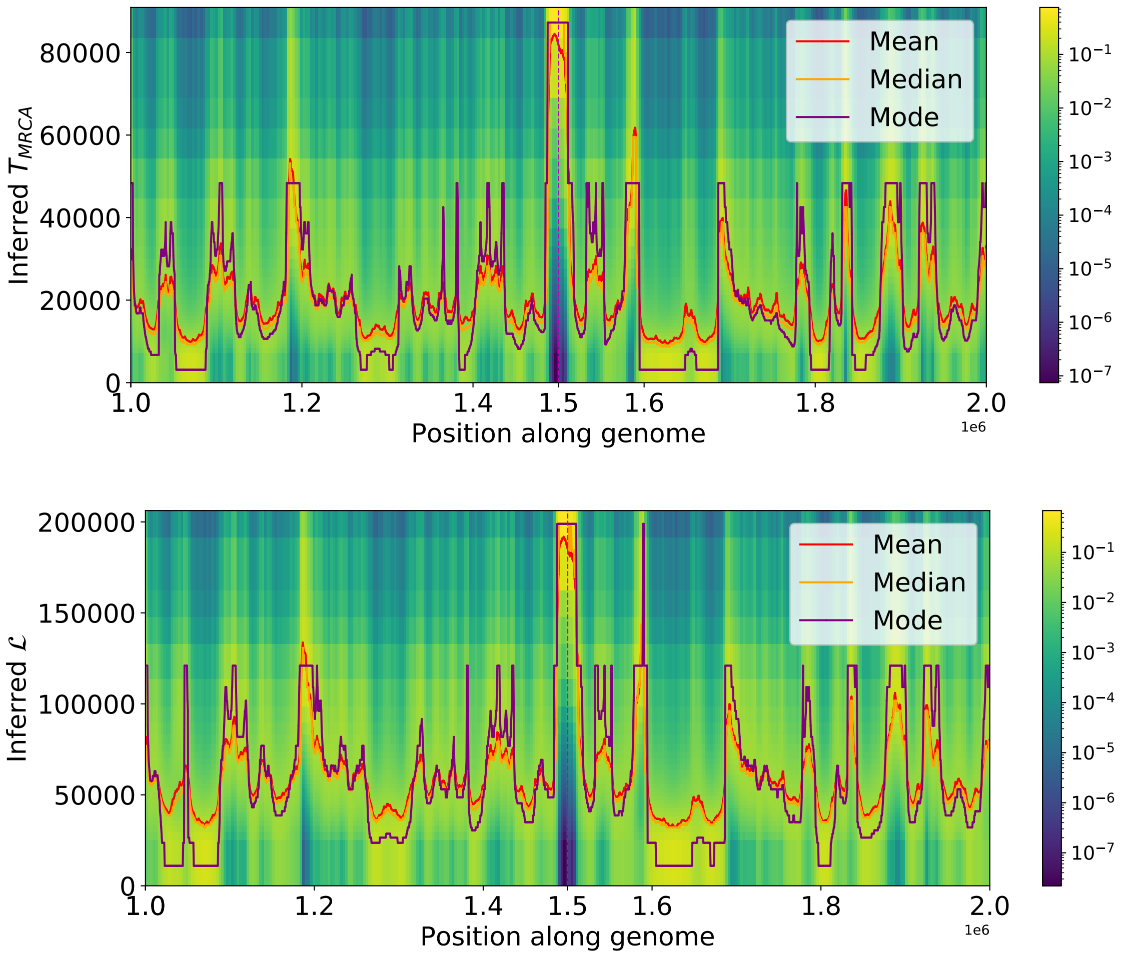


Figure 5.6: Posterior distributions computed by using CHIMP-PD to analyze data collected from a balancing selection simulation with $s = 0.02$. This is computed for both T_{MRCA} (top) and $\mathcal{L}$ (bottom). The plot is centered around the site of the beneficial mutation. The mode, mean, and median of the distribution are also computed and displayed. The regions above $T_{MRCA} \approx 55k$ and $\mathcal{L} \approx 135k$ generations are subdivided into finer states that were included to increase precision in the distant past, and we see that the mode, median, and mean all spike into these states near the selected mutation, showing clear evidence of a signal. At 1.59mbp, we see an example of a spike in the background, though it is highly localized compared to the signal due to the 20kbp region under balancing selection.

long periods of time, and thus span very narrow tracts of the genome. If we simulated selection over a narrow range (or for a single SNP) instead of a 20kbp region, distinguishing true signals from these background fluctuations would likely be more difficult. We show a characteristic example of posterior plots in Figure 5.6), for a dataset with $s = 0.02$. The signal appears to be very sharp, with a very steep drop-off away from the selected region. The suddenness of this drop-off is largely due to the age of the haplotypes under selection, since over long spans of evolutionary time, recombination has many more opportunities to decouple the haplotypes of the selected region from the flanking regions of the genome. The simulated scenarios for $s = 0.01,$ and 0.05 also look quite similar to Figure 5.6), so we omit these plots.

5.5 Signals of Adaptation in Human Data

Having established that `CHIMP-PD` is able to uncover patterns of adaptive genetic variation that are consistent with different modes of selection in simulated data, we turn to analyzing real data. We choose to use data samples collected in the 1000 Genomes Project [42], and focus this study on three populations in particular: GBR (British and Scottish samples), CHB (Han Chinese samples, collected in Beijing), and YRI (Yoruba samples, collected in Ibadan, Nigeria). After partitioning the individuals by their population labels, we used data from 90 GBR individuals, 100 CHB individuals, and 175 YRI individuals, and analyzed these samples using the composite likelihood model where the haplotypes were partitioned into groups of 10, and the log-likelihoods summed at the end. Under this scheme, we used `CHIMP` to infer the demographic history for each population using data across the entire genome (22 chromosomes). The results of this inference are shown in Figure 5.7. The bottleneck at $\approx 100,000$ generations in the past in the GBR and CHB populations is evidence of an out-of-Africa migration, showing general consistency with other results [16] .

After establishing the baseline demographic history using `CHIMP`, we used `CHIMP-PD` to

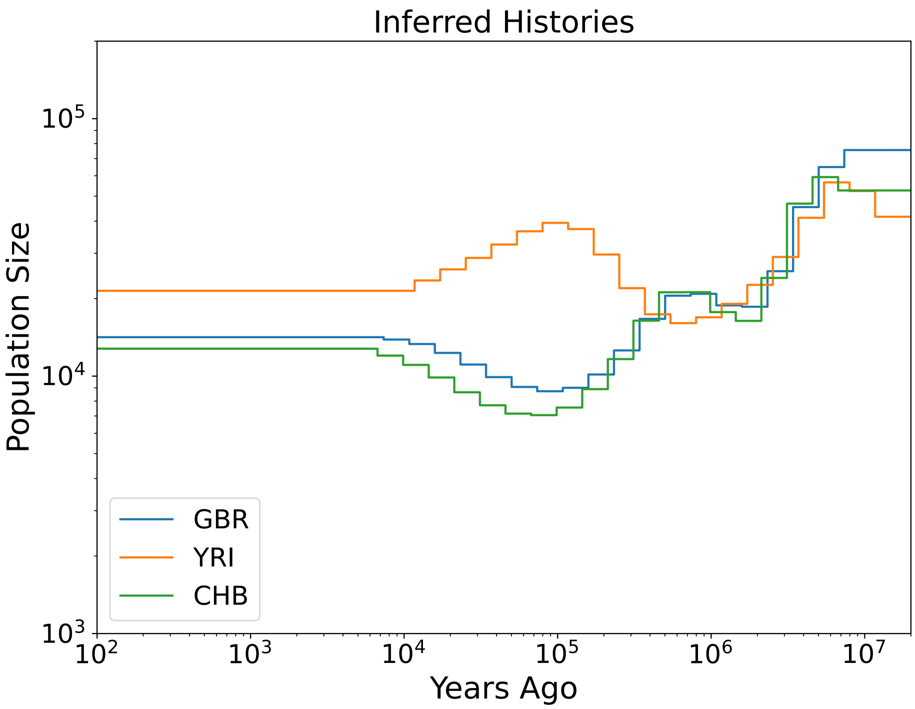


Figure 5.7: Results of demographic inference for GBR, YRI, and CHB. These results were computed with a modest amount of regularization to achieve smoother functions. As expected, we see evidence of a bottleneck for the out-of-Africa migration in the GBR and CHB populations and similar dynamics in the distant past for all populations.

infer the posterior distributions in two specific regions of the genome: the LCT/MCM6 region, and the MHC locus. These regions provided instances where a selective sweep and balancing selection, respectively, have been well established in the literature ([41, 43, 44, 45]), and thus were good candidates to analyze with `CHIMP-PD` to calibrate it to real data.

5.5.1 Studying LCT/MCM6 with `CHIMP-PD`

The LCT gene (on Chromosome 2) contains information to produce the protein lactase, which allows the digestion of milk. All humans produce lactase as infants, and while some humans decrease their production of LPH (an enzyme that helps break down lactase) after 5-10 years of age [46], many do not. It is believed that lactase-persistence is historically linked to substantial fitness benefits, due the domesticization of cattle and the increased

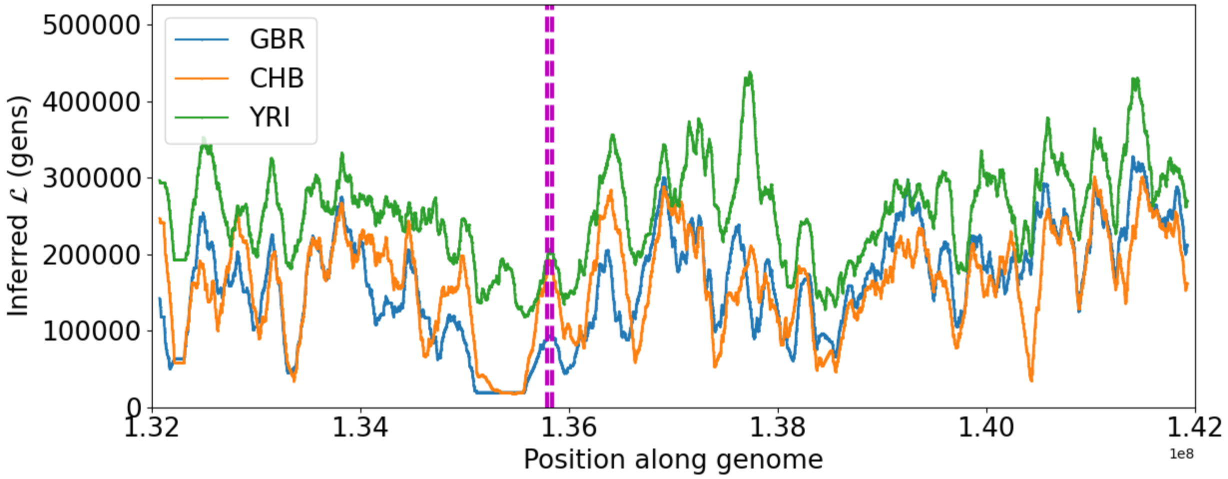


Figure 5.8: $\mathcal{L}$ values near LCT, inferred using CHIMP-PD with the mode of the posterior as the estimator. The purple dashed lines indicate the location of LCT, and near this we see evidence for the very small trees that are consistent with a recent selective sweep. We see strongest evidence of this sweep in the European population GBR, and weaker evidence in CHB, and weakest of all in YRI.

ability to exploit milk as a source of nutrition. Since lactase-persistence is a relatively recent development in our evolutionary history, we are likely to see evidence of selection in the genome [43]. This selection is some of the strongest observable selection in the genome, and the selection coefficient has been estimated to be between 0.02 and 0.05 [41](ie. homozygous presence of the beneficial allele confers a 2-5% fitness advantage). Previous studies have identified multiple SNPs that are linked to lactase-persistence (all in the MCM6 region just upstream of LCT), and while different SNPs appear to be tied to lactase-persistence in different geographic populations, in northern European populations, lactase persistence has been primarily tied to SNP rs4988235 [46], though through population diffusion and migration, the beneficial allele at this location has more recently dispersed to other populations.

For this investigation, we were primarily concerned with the region near aforementioned SNP rs4988235, and sought to corroborate previous studies with CHIMP-PD by demonstrating that there is a strong signal of selection in the posterior distribution of T_{MRCA} and $\mathcal{L}$ in European populations. We analyzed the individuals from each population separately (YRI, CHB, and GBR) assuming recombination and mutation rates $r = \mu = 1.25 \cdot 10^{-8}$ and the demographic histories that were inferred in Figure 5.7. The posterior was computed across chromosome 2 by analyzing the haplotypes in groups of 10 (as described in Section 5.2). For this analysis, the states for the state variable were determined using partitions such that the marginal probability of finding the CHMM in each state was 0.02. Then the lowest state (corresponding to the smallest trees), was broken up further into 5 additional states with marginal probabilities $\{0.002, 0.003, 0.005, 0.005, 0.005\}$. The highest state was also broken up similarly such that the highest resolution of the state was at the tail of the distribution (ie. with marginal probabilities $\{0.005, 0.005, 0.005, 0.003, 0.002\}$). Thus, the CHMM was defined with 58 possible states, with very small marginal probabilities of falling in the lowest or highest state. This is similar to the procedure in Section 5.4, in order to obtain better resolution for very small and large trees, and to accentuate large deviations from the the null-hypothesis of a neutral background.

We show the results of inferring $\mathcal{L}$ in the LCT/MCM6 region in Figure 5.8. The trajectories plotted use the mode of the posterior as the estimate. In the GBR and CHB populations, we see very small trees a little to the left of the LCT gene which indicates a region where the allele (or set of alleles) under selection have swept to high frequencies. This is not observed in the YRI population. These results are consistent with the selection signals detected by other scans that use statistics computed from population genetic data ([47]: fig 1b).

5.5.2 Studying MHC with `CHIMP-PD`

The major histocompatability complex (MHC) is a region of chromosome 6 that contains several polymorphic genes that code for proteins involved in the adaptive immune system. These human leukocyte antigen (HLA) genes are some of the most polymorphic genes that we know of, and their diversity within a population means that most people are heterozygous at HLA alleles. Combined with the fact that expression of these alleles is codominant, this

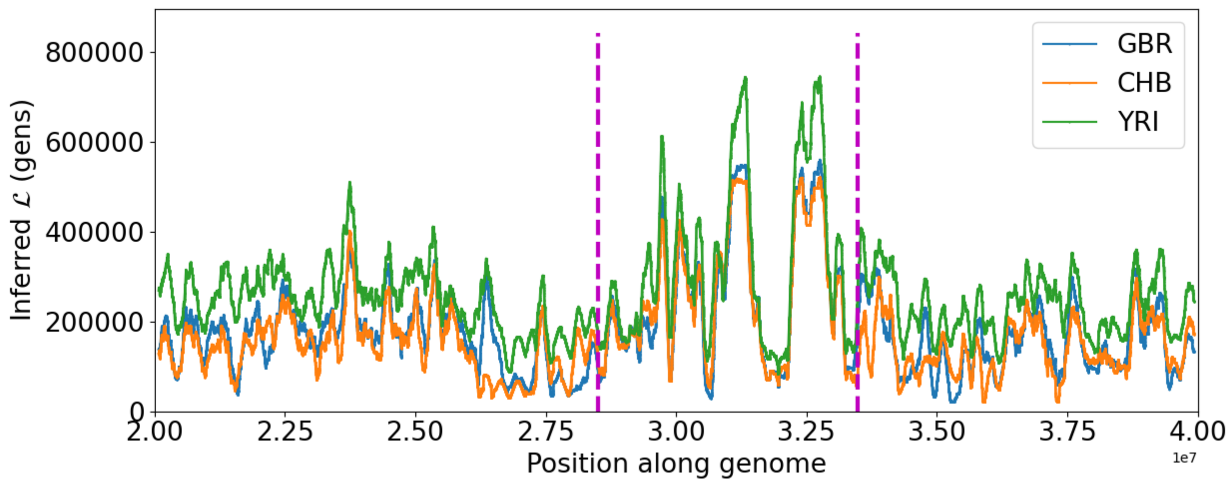


Figure 5.9: $\mathcal{L}$ values near the MHC, inferred using CHIMP-PD with the mode of the posterior as the estimator. The MHC is the region bounded by the purple dashed lines, and within this region we see evidence for very large (deep) trees consistent with the signature for balancing selection in each of the three populations.

allows for each individual to express a great number of HLA molecules, thus aiding in immune function [45]. For the population as a whole, we see that these dynamics make it desirable to maintain a diverse pool of alleles, so competing alleles are retained at high frequency within the broader genetic pool for long periods of time, with little push for one to push the other out, making this an example of balancing selection dynamics [44]. By analyzing the MHC region, we will demonstrate that CHIMP-PD can be used to uncover signals of balancing selection.

Using CHIMP-PD, we computed the posterior distributions for T_{MRCA} and $\mathcal{L}$ for each population separately using the same method as for LCT/MCM6 (including the additional states for very large and small trees). In Figure 5.9, we show the results of this inference. In the MHC region, we see very clear evidence for genes that have very large trees, indicating that the variants observed in each population have to go into the very distant past to find a common ancestor. This is consistent with the hypothesis that this variation has been maintained by balancing variation in each of the populations CHB, GBR, and YRI. The fact that the YRI trees appear larger for these regions under balancing selection also suggests that the YRI population is the oldest (or that the variation in the YRI population is similar to an ancestral pattern of variation), when compared to the GBR and CHB populations, where the bottleneck due to the out-of-Africa migration may have impacted the diversity of SNPs, and leads to smaller trees for the variation observed in modern samples. To sum up, we see that CHIMP-PD is able to uncover such signals of strong balancing selection.

5.6 Conclusion

In this chapter we have begun to explore the possibilities for detecting adaptive variation using CHIMP-PD. We started by testing the ability of CHIMP-PD to accurately infer the T_{MRCA} and $\mathcal{L}$ in neutral simulations, and explored different estimates for the state variable that can be obtained from the posterior distribution that CHIMP-PD returns. We saw that the mean and median estimates performed very well, achieving significantly better accuracy than the estimates obtained from Relate. The mode estimate did not perform as well overall, but was more accurate when the true genealogical trees were very large or small, and thus we believe the mode still has utility in detecting adaptive variation since regions of adaptive variation are often characterized by these outliers.

We then turned to simulations of selective processes, focusing on the archetypes of a selective sweep and balancing selection in particular. We simulated multiple replicates of data for these two scenarios, across a range of selection coefficients, and found that CHIMP-PD has great potential for detecting strong selection even with a simplistic method to classify a region as being under selection. For weaker signals in real data, we will likely need an approach that leverages better signal analysis techniques, and CHIMP-PD can easily be incorporated into such sophisticated pipelines.

Finally, we looked at two regions of the human genome, LCT and the MHC, and analyzed data from the 1000 Genomes Project at these loci. At these locations, we demonstrated that CHIMP-PD is able to replicate the results of other studies, and uncovers strong evidence of a selective sweep near LCT, and balancing selection at the MHC. These preliminary studies suggest that with more sophisticated signal detection criterion, `CHIMP-PD` can be developed into a tool with which the genome can be scanned for patterns consistent with adaptive variation. It can also be as a supplemental tool in other analyses of specific genes to efficiently understand how the local genealogy fits into a broader narrative of gene function.

CHAPTER 6

FINAL THOUGHTS

In this work we have described the potential that CHMMs offer as a tool for analyzing the vast numbers of whole genome sequences that are being collected from across populations. We have developed, bench-marked, and applied the novel CHMM method `CHIMP`. We started by presented two implementations, `CHIMP`-T_{MRCA} and `CHIMP`-$\mathcal{L}$, and using them to perform inference of past population sizes in a single population. The methods use either the T_{MRCA} or the total branch length $\mathcal{L}$ of the local genealogies as the underlying hidden state in this HMM framework. We detailed systems of differential equations derived from the ancestral process that can be used to compute the respective transition and emission probabilities. These differential equations can be computationally intensive to solve, particularly for $\mathcal{L}$, but we present solution schemes that exploit a combination of approximations and exact equations to obtain solutions. Furthermore, the framework presented here can be seen as a generalization of most previous CHMM methods, in that it can readily be modified to use pairwise coalescent times [`PSMC` and `MSMC2`, 13, 15], first-coalescent times [`MSMC`, 14], as well as the coalescent time of a distinguished pair [`SMC++`, 16] as the hidden state.

We applied `CHIMP` for demographic inference from simulated data in a variety of scenarios and compared the results to other state-of-the-art methods, specifically `MSMC2` and `Relate`. While `CHIMP`-$\mathcal{L}$ is intriguing from a theoretical perspective, it currently does not seem suitable for demographic inference since inference is slow and less accurate than `CHIMP`-T_{MRCA}, although more efficient approximations may ameliorate this issue. We observed that `CHIMP`-T_{MRCA} performed comparably to other methods in most scenarios for time-frames more than 500 generations before present and outperformed them for very ancient times beyond 100,000 generations before present. Its runtimes are similar to those of the other methods in the tests we performed, and `CHIMP`-T_{MRCA} scales very well to large samples. `CHIMP` can also be run on unphased data and certain pseudo-haploid datasets, whereas `Relate` requires phased data, and MSMC2 can only be run in a limited capacity without phased data. We believe, that this makes CHIMP-T_{MRCA} a flexible alternative to other methods when analyzing large data sets, especially in scenarios where high quality assessment of haplotype phase is not available, which includes non-model systems were reliable reference panels are not available.

Next we showed how the modeling framework that we developed can be extended beyond inference for a panmictic single-population model. We described CHIMP-S, a method that extends this framework to a population-split model, and we demonstrated inference of demographic parameters using CHIMP-S for simulated data. This highlighted the flexibility of the modeling framework, and we anticipate that it can be adapted to other demographic situations such as those with continuous migration or those that incorporate asynchronous ancient DNA samples.

We then showed how to use CHIMP-PD to obtain the posterior distribution across states, and how estimates for T_{MRCA} and $\mathcal{L}$ can be constructed from these distributions. We benchmarked the accuracy of these estimates (comparing them to Relate). With CHIMP-PD's accurate estimates of the state variable, we showed that signals of selective sweeps and balancing selection can be uncovered in a series of simulation studies. Then we demonstrated the use of CHIMP-PD on human data, focusing on signals of selection near the LCT/MCM6 gene and the MHC.

Taken in total, these various projects highlight the flexibility and utility that CHMMs (and in particular, CHIMP) can provide in aiding genetic studies. They can be used on new genetic data, when little is known about the populations in question beforehand, to infer the demographic history, or scan for selection. They can also be used in a more focused way to understand the variation patterns at specific regions, aiding other studies in trying to understand the evolutionary narrative behind certain genes.

REFERENCES

[1] Nick Barton, Joachim Hermisson, and Magnus Nordborg. Why structure matters. *eLife*, 8:e45380, 2019.

[2] Xiaoming Liu and Yun-Xin Fu. Exploring population size changes using snp frequency spectra. *Nature Genetics*, 47(5):555–559, 05 2015.

[3] Anand Bhaskar, Y. X. Rachel Wang, and Yun S. Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, 25(2):268–279, 01 2015.

[4] Julia A Palacios, Amandine Véber, Lorenzo Cappello, Zhangyuan Wang, John Wakeley, and Sohini Ramachandran. Bayesian estimation of population size changes by sampling tajima's trees. *Genetics*, 213(3):967–986, 2/15/2021 2019.

[5] Alon Keinan and Andrew G Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–3, May 2012.

[6] Sharon R. Browning and Brian L. Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*, 97(3):404–418, 2021/02/14 2015.

[7] Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics*, 91(5):809–22, Nov 2012.

[8] Matthew D. Rasmussen, Melissa J. Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):1–27, 05 2014.

[9] Jerome Kelleher, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers,

and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, 2019.

[10] Leo Speidel, Marie Forest, Sinan Shi, and Simon R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, 2019.

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

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

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

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

[15] Ke Wang, Iain Mathieson, Jared O'Connell, and Stephan Schiffels. Tracking human population structure through time from whole genome sequences. *PLoS Genetics*, 16(3):1–24, 03 2020.

[16] Jonathan Terhorst, John A. Kamm, and Yun S. Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, 2017.

[17] Sara Sheehan, Kelley Harris, and Yun S Song. Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2/15/2021 2013.

[18] Matthias Steinrücken, Jack Kamm, Jeffrey P. Spence, and Yun S. Song. Inference of complex population histories using whole-genome sequences from multiple populations. *Proceedings of the National Academy of Sciences*, 116(34):17115, 08 2019.

[19] Jeffrey P Spence, Matthias Steinrücken, Jonathan Terhorst, and Yun S Song. Inference of population history using coalescent hmms: review and outlook. *Current Opinion in Genetics & Development*, 53:70–76, 2018.

[20] Thibaut Paul Patrick Sellinger, Diala Abu-Awad, and Aurélien Tellier. Limits and convergence properties of the sequentially markovian coalescent. *Molecular Ecology Resources (Early Online)*, 2021/07/02 2021.

[21] Alexey Miroshnikov and Matthias Steinrücken. Computing the joint distribution of the total tree length across loci in populations with variable size. *Theoretical Population Biology*, 118:1–19, 2017.

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

[23] Robert C Griffiths and Paul Marjoram. An ancestral recombination graph. *Progress in Population Genetics and Human Evolution*, 87:257–270, 1997.

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

[25] Jerome Kelleher, Alison M. Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5):1–22, 2016.

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

[27] Paul Marjoram and Jeff D. Wall. Fast "coalescent" simulation. *BMC Genetics*, 7(1):16, 2006.

[28] Richard Durrett. *Probability Models for DNA Sequence Evolution.* Springer, 2008.

[29] Christopher Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[30] John Wakeley. *Coalescent Theory: An Introduction.* Roberts & Co. Publishers, 2009.

[31] J. R. Dormand and P. J. Prince. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980.

[32] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 3/25/2021 1965.

[33] Kris V Parag and Oliver G Pybus. Robust design for coalescent model inference. *Systematic Biology*, 68(5):730–743, 3/29/2021 2019.

[34] Jeffrey R Adrion, Christopher B Cole, Noah Dukler, Jared G Galloway, Ariella L Gladstein, Graham Gower, Christopher C Kyriazis, Aaron P Ragsdale, Georgia Tsambos, Franz Baumdicker, Jedidiah Carlson, Reed A Cartwright, Arun Durvasula, Ilan Gronau, Bernard Y Kim, Patrick McKenzie, Philipp W Messer, Ekaterina Noskova, Diego Ortega-Del Vecchyo, Fernando Racimo, Travis J Struck, Simon Gravel, Ryan N Gutenkunst, Kirk E Lohmueller, Peter L Ralph, Daniel R Schrider, Adam Siepel, Jerome Kelleher, and Andrew D Kern. A community-maintained standard library of population genetic models. *eLife*, 9:e54967, 2020.

[35] Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genetics*, 5(10):1–11, 10 2009.

[36] Julien Jouganous, Will Long, Aaron P. Ragsdale, and Simon Gravel. Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics*, 206(3):1549–1567, 07 2017.

[37] Axel Barlow, Stefanie Hartmann, Javier Gonzalez, Michael Hofreiter, and Johanna L A Paijmans. Consensify: A method for generating pseudohaploid genome sequences from palaeogenomic datasets with reduced error rates. *Genes*, 11(1):50, 01 2020.

[38] Caleb Ki and Jonathan Terhorst. Exact decoding of the sequentially markov coalescent. *bioRxiv*, 01 2020. `https://doi.org/10.1101/2020.09.21.307355`.

[39] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, 2007.

[40] Benjamin C Haller and Philipp W Messer. SLiM 3: Forward genetic simulations beyond the wright–fisher model. *Molecular Biology and Evolution*, 36(3):632–637, jan 2019.

[41] Iain Mathieson. Estimating time-varying selection coefficients 2 from time series data of allele frequencies. *bioRxiv*, 2020.

[42] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, and Richard M. Durbin et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, sep 2015.

[43] Jesper T. Troelsen. Adult-type hypolactasia and regulation of lactase expression. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1723(1):19–32, 2005.

[44] Cock van Oosterhout. A new theory of MHC evolution: beyond selection on the immune genes. *Proceedings of the Royal Society B: Biological Sciences*, 276(1657):657–665, nov 2008.

[45] Walport M et al. Janeway Ca Jr, Travers P. *Immunobiology: The Immune System in Health and Disease*. Garland Science, NY, 5th edition, 2001.

[46] Jesper T Troelsen, Jørgen Olsen, Jette Møller, and Hans SjÖstrÖm. An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology*, 125(6):1686–1694, dec 2003.

[47] Sjödin P. Skoglund P. et al. Schlebusch, C. Stronger signal of recent selection for lactase persistence in maasai than in europeans. *European Journal of Human Genetics*, 2012.