Inferring Population Size Histories using Coalescent Hidden Markov Models with T_{MRCA} and Total Branch Length as Hidden States

Gautam Upadhya[1] and Matthias Steinrücken[2,3]

[1]Department of Physics, University of Chicago, Chicago, IL, USA
[2]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA
[3]Department of Human Genetics, University of Chicago, Chicago, IL, USA

May 22, 2021

Abstract

Unraveling the complex demographic histories of natural populations is a central problem in population genetics. Understanding past demographic events is of general anthropological interest and is also an important step in establishing accurate null models when identifying adaptive or disease-associated genetic variation. An important class of tools for inferring past population size changes are Coalescent Hidden Markov Models (CHMMs). These models make efficient use of the linkage information in population genomic datasets by using the local genealogies relating sampled individuals as latent states that evolve across the chromosome in an HMM framework. Extending these models to large sample sizes is challenging, since the number of possible latent states increases rapidly.

Here, we present our method `CHIMP` (CHMM History-Inference ML Procedure), a novel CHMM method for inferring the size history of a population. It can be applied to large samples (hundreds of haplotypes) and only requires unphased genomes as input. The two implementations of `CHIMP` that we present here use either the height of the genealogical tree (T_{MRCA}) or the total branch length, respectively, as the latent variable at each position in the genome. The requisite transition and emission probabilities are obtained by numerically solving certain systems of differential equations derived from the ancestral process with recombination. The parameters of the population size history are subsequently inferred using an Expectation-Maximization algorithm. In addition, we implement a composite likelihood scheme to allow the method to scale to large sample sizes.

We demonstrate the efficiency and accuracy of our method in a variety of benchmark tests using simulated data and present comparisons to other state-of-the-art methods. Specifically, our implementation using T_{MRCA} as the latent variable shows comparable performance and provides accurate estimates of effective population sizes in intermediate and ancient times. Our method is agnostic to the phasing of the data, which makes it a promising alternative in scenarios where high quality data is not available, and has potential applications for pseudo-haploid data.

1 Introduction

Advances in technology for full-genome sequencing have made it possible to collect large amounts of genomic data for increasingly large samples from many species and diverse population groups. This wealth of genetic data has much to tell us about underlying biological and population genetic phenomena. These datasets can be used to study the relatedness among individuals and understand complex demographic histories, helping unravel population size changes, population structure, and migration events. Moreover, adaptation of beneficial alleles or other forms of selection leave characteristic signatures in genomic sequencing data. In addition, genetic variation across individuals, in human populations in particular, can be used to reveal genetic factors underlying traits relevant for medical and health-related applications. Genome wide association studies are

a widely-used tool to detect such associations. However, effects of population structure can confound the results of these association studies, see for example Barton et al. (2019). Thus, it is imperative to develop population genetic tools for the analysis of whole-genome sequencing data that can infer the underlying demographic history and establish appropriate null models for studying adaptation and associations of genetic variation with certain phenotypes or disease outcomes.

To date, many methods have been presented in the literature that infer different aspects of demographic histories from different signals in the data. The focus in this study is on the inference of the size history of a single population, and here we briefly review methods with similar focus. Several methods perform inference using the site frequency spectrum (SFS), either assuming no linkage between sites (Liu and Fu, 2015; Bhaskar et al., 2015), or complete linkage (Palacios et al., 2019). These methods can be efficiently applied to large sample sizes which particularly improves their ability to infer recent population size changes (Bhaskar et al., 2015; Keinan and Clark, 2012). 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 (Browning and Browning, 2015; Palamara et al., 2012). 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 (Rasmussen et al., 2014; Kelleher et al., 2019; Speidel et al., 2019). 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 McVean and Cardin (2005) based on work by Wiuf and Hein (1999). In this framework, the correlation structure 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-domensional 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` (Li and Durbin, 2011), `MSMC` (Schiffels and Durbin, 2014), `MSMC2` (Wang et al., 2020), `SMC++` (Terhorst et al., 2017), and `diCal` (Sheehan et al., 2013; Steinrücken et al., 2019). 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 be applied to somewhat larger 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 (Spence et al., 2018; Sellinger et al., 2020), but no method performs uniformly well across all parameter regimes.

Here, we present our novel CHMM method, `CHIMP` (**C**HMM **H**istory-**I**nference **M**L **P**rocedure). We present two implementations of `CHIMP` 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. Our method uses the number of derived alleles at a given site as the emission of the HMM, and is therefore agnostic to the phasing of data. Moreover, we implemented a flexible composite likelihood scheme that enables efficient scaling to large sample sizes, resulting in runtime requirements that are comparable to other methods, specifically for the implementation using T_{MRCA}. The latter also shows comparable inference accuracy in intermediate times and outperforms other methods for ancient times. Since the method is agnostic to phasing, it has potential applications to pseudo-haploid data.

The paper is organized as follows. In Section 2, we present the general SMC framework that is the basis for CHMMs and detail the implementation steps for our specific choice of the latent variables. Extending previous work by Miroshnikov and Steinrücken (2017), we present an algorithm to efficiently compute the necessary transition and emission probabilities for the CHMM by numerically solving certain systems of differential equations and incorporate them into a standard Expectation-Maximization (EM) framework for maximum-likelihood inference. In Section 3, we compare the performance of `CHIMP` to other state-of-the-art methods, specifically `MSMC2` and `Relate`, in several simulation studies over a range of demographic scenarios. Lastly, in Section 4, we discuss possible extensions of our framework to infer more complex demographic histories involving multiple populations and to analyze time stratified samples characteristic of ancient DNA datasets. We also discuss how the posterior distribution of the latent states could be applied to study signatures of selection in the genome.

2 Novel CHMM Methods

In this section, we will present relevant background on the Sequentially Markovian Coalescent (SMC) which is the basis for our method, and we will describe our implementation of an HMM framework for inference of past population sizes using T_{MRCA} or the total branch length $\mathcal{L}$ as the hidden states.

2.1 CHMM Model under Variable Population Size

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 of the sample, 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

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

where $N_0 \equiv N(0)$ is an arbitrarily chosen reference population size.

The single-locus coalescent models the genetic variation among n samples at a particular locus in the genome (Kingman, 1982). 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 coalescence is slower when the population size is large. 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 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 (Griffiths and Marjoram, 1997; Hudson, 2002). Specifically, the ARG models the genealogies at each individual locus and their correlations induced by 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 (Hudson, 2002; Kelleher et al., 2016), in many scenarios its applicability in likelihood-based population genetic inference is hindered by its complexity: the space of possible ARGs grows quickly with number of samples and length of genome.

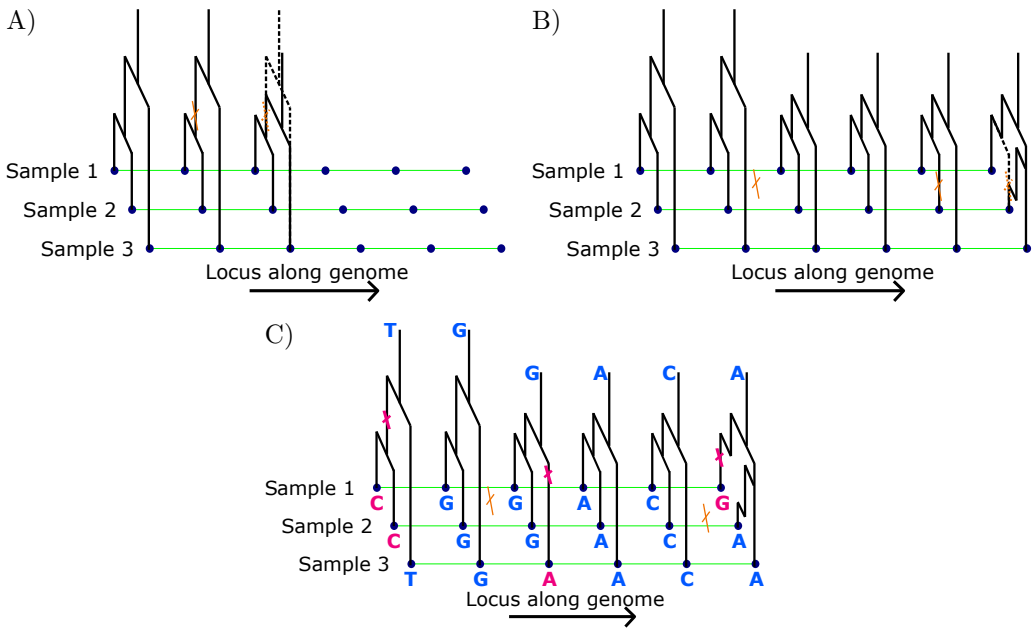


Figure 1: Panel A) shows the marginal genealogy being propagated unchanged along the genome until an ancestral recombination event is encountered, and the genealogy modified accordingly. In panel B), the new genealogy is propagated until a second recombination event is encountered. Panel C) demonstrates a realization of the mutation process along the genealogy at each locus and the resulting observed genetic data.

One factor contributing to the complexity of the ARG is the fact that the marginal genealogies at distant loci can depend on each other (Wiuf and Hein, 1999). The Sequentially Markovian Coalescent (SMC), introduced by McVean and Cardin (2005), 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 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 $\rho = 4N(0)r$, 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, ie. it can coalesce with the regular coalescent rate into the other active 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 1.

Coalescent Hidden Markov models (CHMMs) use the SMC as the basis for computing likelihoods of genomic sequence data. The marginal genealogies are the hidden states in the HMM, and by superimposing mutations onto these trees, 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 1). However, implementing the full model depicted in Figure 1 in a likelihood-based inference method is intractable due to the prohibitively large hidden state space, which is a consequence of 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

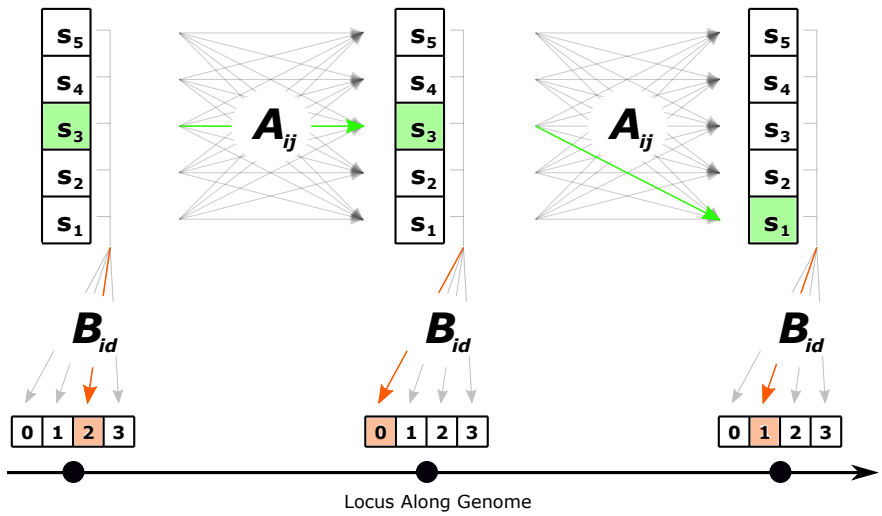


Figure 2: Schematic of our CHMM for a sample of size 4. Information about the underlying tree at each locus is captured by the state s_i. $\mathcal{S} = \{s_1, \ldots, s_5\}$ is the set of intervals into which the respective summary of the tree (T_{MRCA} or $\mathcal{L}$) can fall. The states change 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 $\mathbf{B}$.

for the HMM.

We introduce a novel method, `CHIMP`, which is a CHMM with a one-dimensional hidden state. In our method, 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, the model 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 given by

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

$$B_{id} = \mathbb{P}[d^\ell = d | s^\ell = s_i], \text{ and} \tag{2}$$

$$\Pi_i = \mathbb{P}[s^\ell = s_i]. \tag{3}$$

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 depicts a schematic of the transition and emissions in this CHMM.

2.2 T_{MRCA} as Hidden State

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 2.1. We now describe the numerical

methods used to compute the corresponding transition and emission probabilities in equation (1), (2), and (3). 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 or 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.

2.2.1 Transition Probabilities

To compute the transition probabilities, we employ an augmented ancestral process with recombination, $\mathcal{A}^\rho$, introduced by Miroshnikov and Steinrücken (2017). This process closely resembles the regular ancestral process with recombination described by Simonsen and Churchill (1997) 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. The two decoupled lineages then evolve 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 tuples describing the configuration of lineages, (k_{ab}, k_a, k_b, r). 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 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 (Marjoram and Wall, 2006), which is a more faithful model of the true genetic process than the SMC formulation presented in McVean and Cardin (2005).

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 motivated by the assumption that the recombination rate between two adjacent loci is low, so we expect to see at most one ancestral recombination event separating the genealogies at two neighboring loci. 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 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 by one the number of lineages ancestral to only a, which can either be a coalescent event among the a lineages, or a

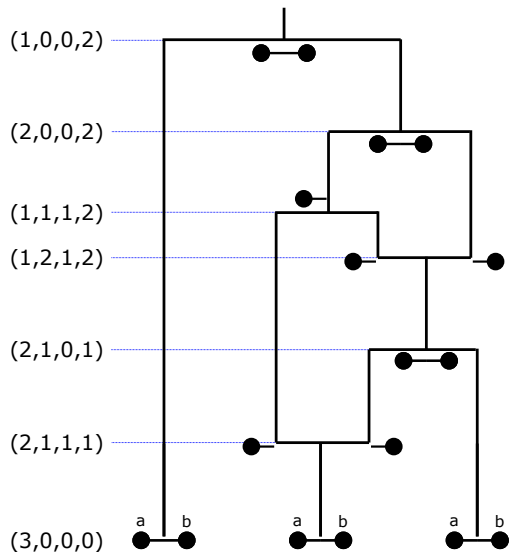


Figure 3: Example trajectory of $\mathcal{A}^\rho(t)$ with the state denoted by tuples. At $t = 0$ there are three lineages of type k_{ab} ancestral to both loci for their respective samples. The lineages split at ancestral recombination events and combine at coalescence events where they find a common 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.

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, 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 $g^\rho_\sigma(t) := \mathbb{P}[\mathcal{A}^\rho(t) = \sigma]$ 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 $\vec{g}^\rho(t) = \left(g^\rho_\sigma(t)\right)_{\sigma\in\mathcal{R}}$ 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)

$$\frac{d}{dt}\vec{g}^\rho(t) = \vec{g}^\rho(t) \cdot \mathbf{Q}^\rho(t), \tag{4}$$

where $\mathbf{Q}^\rho(t)$ is the rate matrix consisting of the rates given in Table 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 (4).

Moreover, from the distribution $\vec{g}(t)$, we can compute the cumulative joint distribution function (CDF) of the T_{MRCA} at the two loci, $\mathbb{P}[T_a \le \tau_a; T_b \le \tau_b]$, where $0 \le \tau_a, \tau_b < \infty$, and T_a and T_b are the T_{MRCA}'s at

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 1: The table shows the possible transitions out of 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.

a and b respectively. Without loss of generality, we assume that $\tau_a < \tau_b$ and obtain

$$\begin{aligned}
&\mathbb{P}[T_a \leq \tau_a; T_b \leq \tau_b] \\
&\quad = \mathbb{P}[T_a \leq \tau_a; T_b \leq \tau_a] + \mathbb{P}\big[T_a \leq \tau_a; T_b \in (\tau_a, \tau_b]\big] \\
&\quad = g_{(1,0,0,0)}(\tau_a) + g_{(1,0,0,1)}(\tau_a) + \mathbb{P}\big[T_a \leq \tau_a; T_b \in (\tau_a, \tau_b]\big] \\
&\quad = g_{(1,0,0,0)}(\tau_a) + g_{(1,0,0,1)}(\tau_a) + g_{(1,0,1,1)}(\tau_a) \cdot \mathbb{P}\big[\sigma(\tau_b) = (1,0,0,1) \mid \sigma(\tau_a) = (1,0,1,1)\big] \\
&\quad = g_{(1,0,0,0)}(\tau_a) + g_{(1,0,0,1)}(\tau_a) + g_{(1,0,1,1)}(\tau_a) \cdot \big[1 - e^{-\int_{\tau_a}^{\tau_b} \lambda(t)dt}\big]. \qquad (5)
\end{aligned}$$

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 (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:

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

$$A_{ij}^{PDF} := A_{ij}^{CDF} - A_{i-1,j}^{CDF} - A_{i,j-1}^{CDF} + A_{i-1,j-1}^{CDF} \qquad (6)$$

$$A_{ij} := \frac{A_{ij}^{PDF}}{\sum_{k \in \mathcal{S}} A_{ik}^{PDF}} \qquad (7)$$

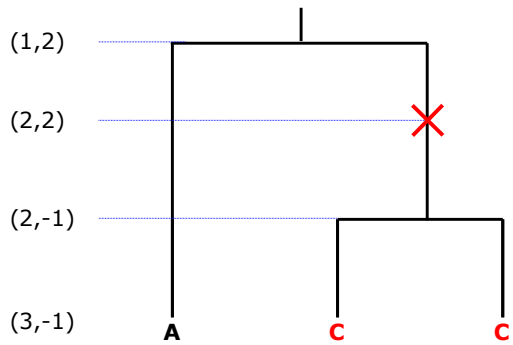


Figure 4: Example trajectory of the ancestral process with mutation for $n = 3$ samples with the state (k, k^*) indicated on the left. The mutation process is superimposed onto the regular genealogical process. In this example, the mutation happens when there are two ancestral lineages, resulting in two samples carrying the derived allele.

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

$$\Pi_i = \sum_{j \in \mathcal{S}} A_{ij}^{PDF}. \tag{8}$$

2.2.2 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 (Griffiths and Marjoram, 1997) 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 and an example trajectory is shown in Figure 4. The possible transitions are either two lineages coalescing 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 $k^* \in \{-1, 2, \ldots, n\}$. 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}.

Similar to the procedure for $\mathcal{A}^\rho$, we collect all transition rates in a matrix $\mathbf{Q}^\theta(t)$. We further define $g_\sigma^\theta(t) := \mathbb{P}[\mathcal{A}^\theta(t) = \sigma]$ 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 $\vec{g}^\theta(t) = \left(g_\sigma^\theta(t)\right)_{\sigma \in \mathcal{M}}$ 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.

Using the distribution of this process, we can compute the cumulative distribution of T_{MRCA} jointly with

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

Table 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.

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

$$\begin{aligned}
&\mathbb{P}[T_{MRCA} \leq \tau; d \text{ derived alleles}] \\
&\quad = \mathbb{P}[\mathcal{A}^\theta(\tau) = (1, k^*) \text{ for } k^* \in \{2, \ldots, n\}; d \text{ derived alleles}] \\
&\quad = \sum_{k^* \in \{2,\ldots,n\}} g^\theta_{(1,k^*)}(\tau) \cdot \mathbb{P}\big[d \text{ derived alleles} | \text{mutation while } k^* \text{ lineages}\big] \\
&\quad = \sum_{k^* \in \{2,\ldots,n\}} g^\theta_{(1,k^*)}(\tau) \cdot \frac{\binom{n-d-1}{k^*-2}}{\binom{n-1}{k^*-1}} \qquad (9)
\end{aligned}$$

for $d \in \{1, \ldots, n-1\}$. The first equality in equation (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. Durrett, 2008, 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:

$$B^{CDF}_{id} := \mathbb{P}[T_{MRCA} \leq t_i; y = d] \qquad (10)$$

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

$$B_{id} := \frac{B^{PDF}_{id}}{\sum_k B^{PDF}_{ik}}. \qquad (11)$$

The transition and emission probabilities can then be used to compute likelihoods of observed sequence data and perform inference using an EM algorithm, which will be described in more detail in section 2.4.

2.3 Total Branch Length as Hidden State

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 outlined in Miroshnikov and Steinrücken (2017) 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.

2.3.1 Transition Probabilities

We use an approach introduced by Miroshnikov and Steinrücken (2017) 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 2.2.1) to compute the requisite distributions for T_{MRCA} and

$$v^\ell(k_{ab}, k_a, k_b, r) := (k_{ab} + k_\ell)\mathbb{1}_{\{k_{ab}+k_\ell>1\}}, \tag{12}$$

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

$$L_\ell(t) := \int_0^t A_\ell(s)ds. \tag{13}$$

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 (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 (12) signifies that we stop accumulating tree length once only a single lineage is left, ie. the T_{MRCA} is reached. Using this notation, we now define the probabilities

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

similar to Miroshnikov and Steinrücken (2017), 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.

Miroshnikov and Steinrücken (2017) show that the values of $\vec{F} := \big(F_\sigma(t, x, y)\big)_{\sigma\in\mathcal{R}}$ can be obtained from the system of PDEs

$$\partial_t\vec{F} + \partial_x\vec{F} \cdot \mathbf{V}^a + \partial_y\vec{F} \cdot \mathbf{V}^b = \vec{F} \cdot \mathbf{Q}^\rho \tag{14}$$

and its corresponding boundary conditions

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

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

$$\mathbf{V}^\ell := \text{diag}\{v^\ell(\sigma)\mathbb{1}_{\{v^\ell(\sigma)>1\}}\},$$

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

As shown by Miroshnikov and Steinrücken (2017), 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

$$\partial_t\vec{F} + \partial_x\vec{F} \cdot \mathbf{V}^a = \vec{F} \cdot \mathbf{Q}^\rho \tag{15}$$

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 introduced in (Miroshnikov and Steinrücken, 2017, Appendix B) to compute the solutions to these PDEs and provide the details of our implementation in Appendix A.

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 is given by

$$\mathbb{P}[\mathcal{L}_a \leq x, \mathcal{L}_b \leq y] = \big[F_{(1,0,0,0)}(t,x,y) + F_{(1,0,0,1)}(t,x,y)\big]\Big|_{t=\frac{max(x,y)}{2}}, \tag{16}$$

where $\mathcal{L}_a$ and $\mathcal{L}_b$ denote the total branch length of the genealogies at locus a and b respectively. 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 (6) and (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 (8).

2.3.2 Emission Probabilities

Computing the emission probabilities closely follows the steps for the transition probabilities outlined in the previous Section 2.3.1. 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 (13) as

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

where $v^\theta(k,k^*) = k$. Similar to Section 2.3.1 and Miroshnikov and Steinrücken (2017), the vector of these probabilities can be obtained as the solution to the following system of PDEs

$$\partial_t \vec{F} + \partial_x \vec{F} \cdot \mathbf{V}^\theta = \vec{F} \cdot \vec{Q}^\theta, \tag{17}$$

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

$$\mathbf{V}^\theta := \text{diag}\{v^\theta(\sigma)\mathbb{1}_{v^\ell(\sigma)>1}\}.$$

The solution to this system can be obtained using a similar approach to Miroshnikov and Steinrücken (2017), and we provide details in Appendix A. Similar to equation (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 the discretization points $x \in \mathcal{S} = \{t_0, t_1, \ldots, t_S\}$ to obtain the entries of the matrix of cumulative probabilities $\mathbf{B}^{CDF}$, which can be substituted into equation (10) and (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.

2.4 Inferring Model Parameters

In this section we detail a procedure for inferring demographic model parameters using the HMM framework introduced in the previous section and introduce some extensions of the algorithm.

2.4.1 Expectation-Maximization Algorithm

We use the Expectation-Maximization (EM) algorithm for HHMs to iteratively infer $\vec{\lambda}$, the parameters of the coalescent rate function $\lambda(t)$ (and consequently the population size history $\eta(t)$). For $\lambda(t)$, we use either a piecewise constant parametrization or a spline parametrization as described in more detail in Appendix B.1, and thus we infer a finite number of parameters. We choose a discretization for the hidden states (independent from the population size history) such that the marginal distribution over the hidden states under a constant population size is approximately uniform (Appendix C).

We denote the parameters in the k-th iteration by $\vec{\lambda}^k$. The initial parameters $\vec{\lambda}^0$ can be specified either by the user or by Watterson's estimator (Appendix B.2). 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 $\vec{\lambda}^k$. In the case of T_{MRCA} as the hidden state, we use a Dormand-Prince algorithm of order 8(5,3) (Dormand and Prince, 1980), to solve the respective ODEs. When using $\mathcal{L}$ as the hidden state, we compute the probabilities by solving the associated PDEs using the scheme detailed in Appendix A, where we again use the Dormand-Prince method for the boundaries that require solving ODEs.

Using these probabilities, we then apply the Forward-Backward algorithm (Bishop, 2006, 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. In Appendix E, we explain how we process input from `vcf`-files to obtain this vector of derived allele counts. The Forward-Backward algorithm yields the likelihood of the current demographic parameters, as well as $\mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \to j) \text{ transitions}]$, $\mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \downarrow d) \text{ emissions}]$, and $\mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[s^1 = i]$, the number of expected transitions from state i to j, expected emissions of d derived alleles given state i, and the expected initial state, respectively, all conditional on the current parameters and the data. We use the scaled implementation (Bishop, 2006, Ch. 13.2.4) for numerical stability. Evaluating the Forward-Backward algorithm at each locus in the genome can become prohibitive. We thus detail two strategies to speed-up these computations in Appendix D: a locus-skipping method that compresses the computations between segregating sites, and a meta-locus method that groups tracts of the genome into meta-loci to reduce the effective number of loci.

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 $\vec{\lambda}^k$,

$$\begin{aligned} Q(\vec{\lambda}|\vec{d},\vec{\lambda}^k) = &\sum_{i\in\mathcal{S}} \log\left(\Pi_i(\vec{\lambda})\right) \cdot \mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[s^1 = i] \\ &+ \sum_{i,j\in\mathcal{S}} \log\left(A_{ij}(\vec{\lambda})\right) \cdot \mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \to j) \text{ transitions}] \\ &+ \sum_{i\in\mathcal{S},d\in\mathcal{D}} \log\left(B_{id}(\vec{\lambda})\right) \cdot \mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \downarrow d) \text{ emissions}], \end{aligned} \tag{18}$$

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

and thus yield the updated parameters for the next iteration are given by

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

Since CHIMP evaluates the Q function numerically, we use the Nelder-Mead simplex optimization procedure for numerical optimization (Nelder and Mead, 1965). We observed some dependence of the results on the shape and orientation of the intial simplex for the Nelder-Mead procedure in each M-step, as the overall search direction can be biased by the orientation. To alleviate this bias, we initialize each M-step with an equilateral simplex centered on the optimal parameters from the previous iteration of the EM algorithm. 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 (Parag and Pybus, 2019) and also has the benefit that the coalescent rates are by design positive. In Appendix B.3, we provide details on different implementations to possibly regularize the population size function in the inference, however, no regularization was used for the simulation studies presented in Section 3. After finding the optimal coalescent rates using the EM algorithm, we invert and scale them to recover the estimates for the population size history $N(k)$.

2.4.2 Composite Likelihood

Our method takes as input 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 more computationally expensive for larger sample sizes. Thus, to efficiently analyze large samples, we modify the EM algorithm to 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 E-step is then performed for each sub-group independently to obtain $\mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \to j) \text{ transitions}]$, $\mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[\#(i \downarrow d) \text{ emissions}]$, and $\mathbb{E}_{\vec{s}|\vec{d},\vec{\lambda}^k}[s^1 = i]$, where $d \in \{0, \ldots, n_s\}$. These expectations are then added across sub-groups to obtain composite expectations for the whole sample. The M-step is subsequently performed jointly for all the sub-groups by using these composite expectations in the objective function defined in equation (18). This composite likelihood scheme is also used to perform parameter inference using sequence data from different chromosomes simultaneously. Thus, holding n_s constant, the E-step scales linearly with the number of samples (parallelizing the independent E-steps could even make the E-step constant in the sample size, though we have not implemented this). Moreover, the computational time of the M-step depends only on n_s, which is especially beneficial if the M-step is computationally more expensive (which we found to be the case when using $\mathcal{L}$ as the hidden state). This composite likelihood scheme is closely related to MSMC2 (Wang et al., 2020), as this method uses all overlapping sub-groups of size $n_s = 2$ for the E-Step, and combines them in a similar way for the M-Step.

3 Results

To evaluate the accuracy of our method, we performed a series of simulation studies on data generated under various demographic scenarios. We 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 (v2.1.2) and Relate (v1.1.3). For each study we used the specified model of the demographic history to simulate $m = 16$ replicates of data using msprime (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 note here that the performance of Relate improved when we simulated and analyzed data with a human recombination map (see Appendix G and Figure G.2). 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 proceeded with a uniform recombination map in our simulation studies.

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 2.4.2, 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 intialize the iterative inference methods, we chose Watterson's estimator for `CHIMP`, as detailed in Appendix B.2, 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 in generation k, or mean signed error, across the replicates

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

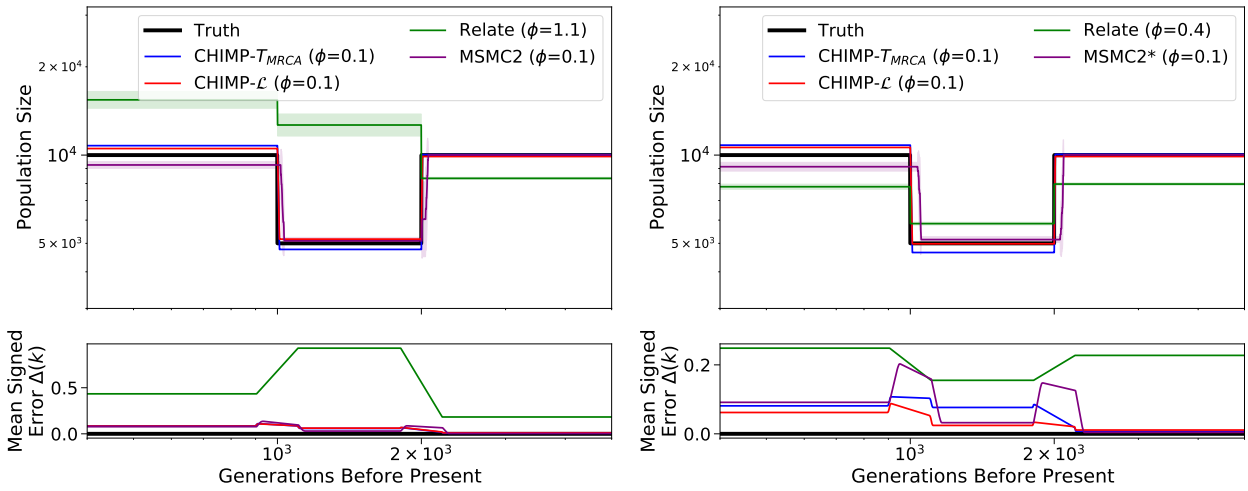
where $\hat{N}^{(j)}$ is the population size estimated in replicate j, and for each method, compute the integral of this quantity $\phi = \int_{k_{\min}}^{k_{\max}} \Delta(k) \frac{1}{k}\, dk$ as a measure of discrepancy from the truth over the full history. Here $k_{\min}$ and $k_{\max}$ are the minimum 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.

3.1 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 2000 generations before present, the population size is reduced to 5,000, but recovers to 10,000 at 1000 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 5 shows the results of the inference using the different methods in the bottleneck 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.

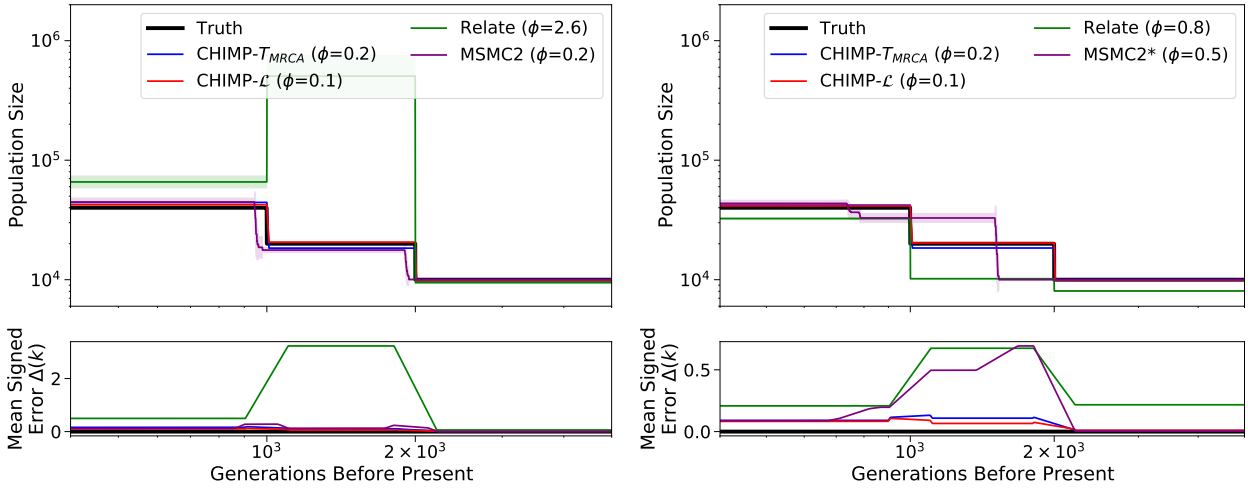
The results for the inference in the piecewise growth scenario are depicted in Figure 6. In this scenario, `CHIMP`-T_{MRCA} and `CHIMP`-$\mathcal{L}$ are again able to infer the population size history with high accuracy. For a



(a) Sample size 10.

(b) Sample size 200.

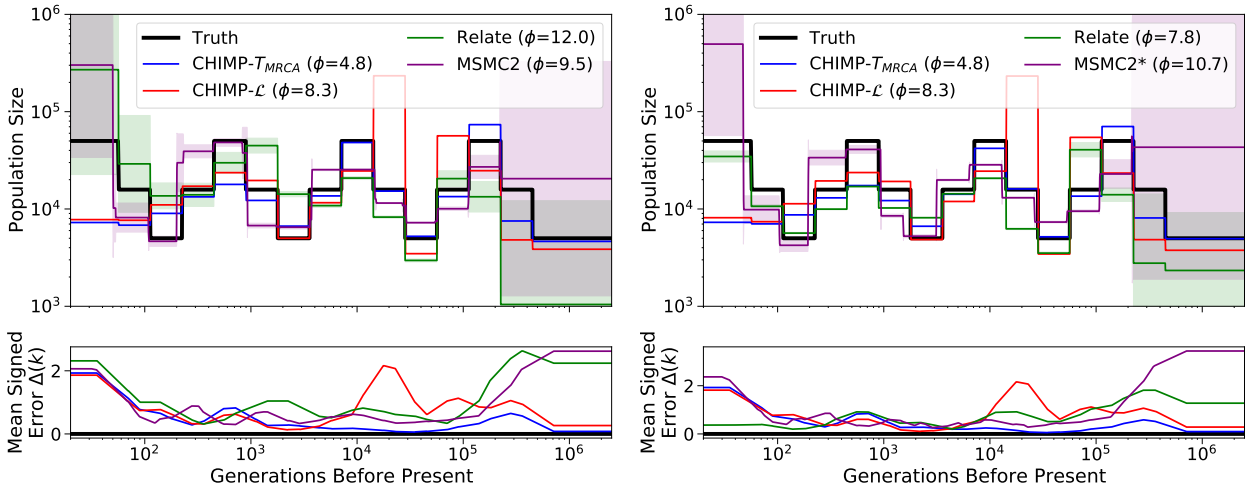
Figure 5: Results of inference in the bottleneck scenario for sample size (a) 10 and (b) 200. We compare the results of `CHIMP`, `MSMC2`, and `Relate` to infer the three population sizes, fixing the change points to match the truth (shown in black). Solid lines are averages over 16 replicates and shaded 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.



(a) Sample size 10.

(b) Sample size 200.

Figure 6: 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.



(a) Sample size 10.

(b) Sample size 200.

Figure 7: 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.

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 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 (Schiffels and Durbin, 2014; Terhorst et al., 2017; Speidel et al., 2019; Sellinger et al., 2020). We will analyze a continuous version of this scenario in Section 3.3, 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 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 7. 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. It smooths the history in the very recent intervals. The estimates are accurate even in the very distant past. 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 3.1) 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.

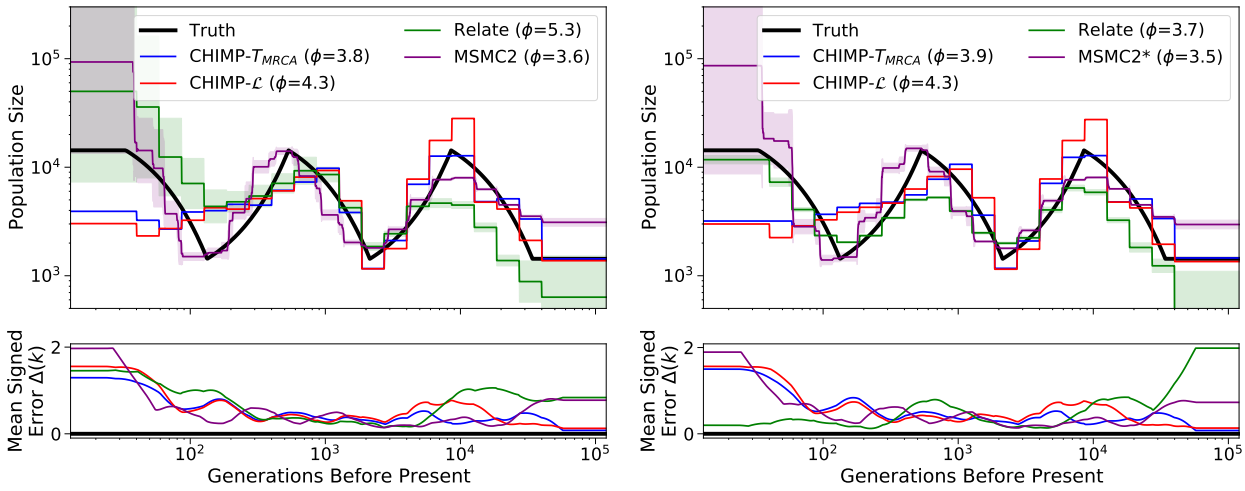
MSMC2 shows accurate performance for intermediate times despite smoothing out some of the oscillations, but demonstrates high variability and a systematic upward bias below 100 generations and above 100,000 generations before present. Its accuracy does not change much between analyzing samples of different 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.

3.3 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 (Adrion et al., 2020, *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 Schiffels and Durbin (2014) and Terhorst et al. (2017). 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 (Gutenkunst et al., 2009; Jouganous et al., 2017). 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 8. 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 again 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 9 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



(a) Sample size 10.

(b) Sample size 200.

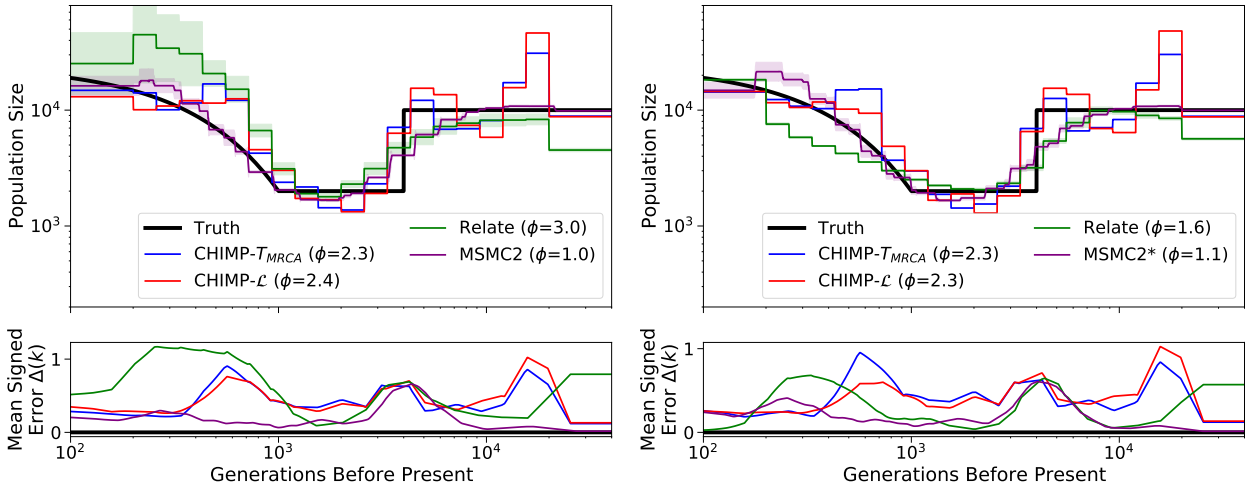
Figure 8: Results of inference in the continuous *sawtooth* scenario for sample size (a) 10 and (b) 200. We compare the results of `CHIMP`, `MSMC2`, and `Relate` using a piecewise constant population size history with 19 change points. Truth shown in black. Solid lines are averages over 16 replicates 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`-$\mathcal{L}$, 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.

3.4 Computational Efficiency

In general, the runtime of the CHMM methods `CHIMP` and `MSMC2` scale linearly with number of loci L. In addition, the `CHIMP` methods scale linearly with the sample size, where we can use the composite likelihood framework introduced in Section 2.4.2 to reduce the effective sample size used in the computation of the transition and emission probabilities. If all pairs of samples are used in `MSMC2`, the method scales quadratically with the sample size, but when using 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. 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`



(a) Sample size 10.

(b) Sample size 200.

Figure 9: 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.

	`CHIMP`-T_{MRCA}	`CHIMP`-$\mathcal{L}$	`MSMC2`*	`Relate`
Bottleneck ($n = 10$, Fig. 5a)	0.1	1.3	4.9	0.1
Bottleneck ($n = 200$, Fig. 5b)	1.2	2.4	5.4	7.0
Piecewise Growth ($n = 10$, Fig. 6a)	0.1	1.7	2.1	0.1
Piecewise Growth ($n = 200$, Fig. 6b)	1.1	2.9	9.7	9.2
Piecewise Sawtooth ($n = 10$, Fig. 7a)	0.9	29.7	4.3	0.2
Piecewise Sawtooth ($n = 200$, Fig. 7b)	7.2	35.2	6.2	8.9
Sawtooth ($n = 10$, Fig. 8a)	1.2	29.9	4.1	0.1
Sawtooth ($n = 200$, Fig. 8b)	7.6	37.9	**5.0	3.0
Bottleneck + Growth ($n = 10$, Fig. 9a)	1.2	28.3	5.7	0.1
Bottleneck + Growth ($n = 200$, Fig. 9b)	7.5	35.1	3.2	3.7

Table 3: 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.

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}.

3.5 Analyzing Unphased and Pseudo-haploid Data

Our method `CHIMP` can 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 (Barlow et al., 2020). 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 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). 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 7a. 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. We note though that our method still relies on the information that no segregating sites are observed between the SNPs, which might not be available in low coverage sequencing data. However, we do believe that the capability to analyze pseudo-haploid data generated in this way presents an exciting avenue for future extensions.

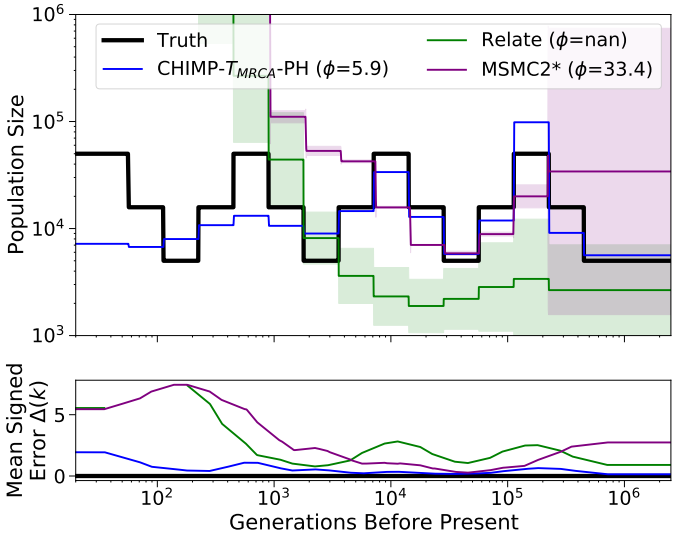


Figure 10: Results of inference using 10 pseudo-haploids simulated under the piecewise sawtooth demography. We compare the results of CHIMP $- T_{MRCA}$ using the pseudo-haploid option, MSMC2, and Relate, fixing the change points to match the truth (shown in black). Solid lines are averages over 16 replicates 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.

3.6 Summary

Table 4 shows a summary of the performace 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-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 $12,000$.

In summary, 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

	Gener. bp	CHIMP-T_{MRCA}	CHIMP-$\mathcal{L}$	MSMC2	Relate
Sample size		Any	Any	Limited	Any
Parametrization		Flexible	Flexible	Limited	Flexible
Accuracy	Few parameters	High	High	High	Low
	$k < 100$	Low	Low	Low	High (large n)
	$100 < k < 500$	Low	Low	High	High (large n)
	$500 < k < 100,000$	High	Mixed	High	Mixed
	$k > 100,000$	High	High	Low	Low
Runtime		Fast	Slow	Fast (limited n)	Fast
Unphased data		Yes	Yes	Possible	No
Pseudo-haploid		Limited	Limited	No	No

Table 4: Summary of the performance and features of the different methods compared in our simulation study. The ranges are given in generations before present.

useful alternative to other existing methods in situations where high quality data is not available. Overall, the inference accuracy and runtime 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.

4 Discussion

Here, we presented two novel flexible Coalescent HMM methods, CHIMP-T_{MRCA} and CHIMP-$\mathcal{L}$, 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, Li and Durbin, 2011; Wang et al., 2020), first-coalescent times (MSMC, Schiffels and Durbin, 2014), as well as the coalescent time of a distinguished pair (SMC++, Terhorst et al., 2017) 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. However, 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.

We note that, when performing the analyses of simulated data presented here and designing comparisons between the methods that can be deemed fair, the flexibility of the user-interface and the heuristics to choose an a-priori discretization of the population size history for inference impacts the applicability and performance of the different methods. We showcase this with an application of the methods to the bottleneck followed by growth scenario in Figure F.1, where we ran all methods using their default parameters. In Appendix F, we also detail the heuristic we implemented in CHIMP for determining a default parameterization. We found 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 Ki and Terhorst (2020), present interesting alternatives.

The modeling framework developed here has potential applications beyond inferring the population size history of a single population. The differential equations that we presented to compute the requisite probabilities for the CHMM were derived from the ancestral process in a panmictic population, but the approach can be extended to structured populations by augmenting the state space, enabling inference of migration rates and divergence times, and characterizing admixture between populations. Moreover, samples at different points in time (e.g. ancient samples), can be readily incorporated into the ancestral process and thus the inference framework as well. In addition, using the model and possible extensions presented here to characterize the posterior distribution of the local genealogies has many interesting applications. Different forms of selection impact the local genealogies around beneficial alleles, and thus the posterior distribution of the T_{MRCA} or the total branch length $\mathcal{L}$ can help identifying and characterizing adaptive genetic variation (Stern et al., 2019; Speidel et al., 2019).

Software availability

Our software CHIMP (**C**HMM **H**istory-**I**nference **M**L **P**rocedure) is available for download at `https://www.github.com/steinrue/chimp`. Instruction on how to run the software and scripts to recreate the simulation study presented here can be found at this address as well.

Acknowledgements

We thank the Steinrücken, Novembre and Berg lab for many inputs and helpful feedback. We also thank Maryn Carlson and Arjun Biddanda for feedback on the manuscript. In addition, we would like to thank Margarita Orlova for assistance in performing the simulation studies.

References

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-Del Vecchyo, D., Racimo, F., Struck, T. J., Gravel, S., Gutenkunst, R. N., Lohmueller, K. E., Ralph, P. L., Schrider, D. R., Siepel, A., Kelleher, J., and Kern, A. D. (2020). A community-maintained standard library of population genetic models. *eLife*, **9**, e54967.

Barlow, A., Hartmann, S., Gonzalez, J., Hofreiter, M., and Paijmans, J. L. A. (2020). Consensify: A method for generating pseudohaploid genome sequences from palaeogenomic datasets with reduced error rates. *Genes*, **11**(1), 50.

Barton, N., Hermisson, J., and Nordborg, M. (2019). Why structure matters. *eLife*, **8**, e45380.

Bhaskar, A., Wang, Y. X. R., and Song, Y. S. (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Browning, S. R. and Browning, B. L. (2015). 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.

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

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. Springer.

Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. *Progress in Population Genetics and Human Evolution*, **87**, 257–270.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genetics*, **5**(10), 1–11.

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

Jouganous, J., Long, W., Ragsdale, A. P., and Gravel, S. (2017). Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics*, **206**(3), 1549–1567.

Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, **336**(6082), 740–3.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, **12**(5), 1–22.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, **51**(9), 1330–1338.

Ki, C. and Terhorst, J. (2020). Exact decoding of the sequentially markov coalescent. *bioRxiv*. `https://doi.org/10.1101/2020.09.21.307355`.

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

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

Liu, X. and Fu, Y.-X. (2015). Exploring population size changes using snp frequency spectra. *Nature Genetics*, **47**(5), 555–559.

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

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

Miroshnikov, A. and Steinrücken, M. (2017). Computing the joint distribution of the total tree length across loci in populations with variable size. *Theoretical Population Biology*, **118**, 1–19.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, **7**(4), 308–313.

Palacios, J. A., Véber, A., Cappello, L., Wang, Z., Wakeley, J., and Ramachandran, S. (2019). Bayesian estimation of population size changes by sampling tajima's trees. *Genetics*, **213**(3), 967–986.

Palamara, P. F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet*, **91**(5), 809–22.

Parag, K. V. and Pybus, O. G. (2019). Robust design for coalescent model inference. *Systematic Biology*, **68**(5), 730–743.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, **10**(5), 1–27.

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

Sellinger, T., Awad, D. A., and Tellier, A. (2020). Limits and convergence properties of the sequentially markovian coalescent. *bioRxiv*. `https://doi.org/10.1101/2020.07.23.217091`.

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

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

Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, **51**(9), 1321–1329.

Spence, J. P., Steinrücken, M., Terhorst, J., and Song, Y. S. (2018). Inference of population history using coalescent hmms: review and outlook. *Current Opinion in Genetics & Development*, **53**, 70–76.

Steinrücken, M., Kamm, J., Spence, J. P., and Song, Y. S. (2019). Inference of complex population histories using whole-genome sequences from multiple populations. *Proceedings of the National Academy of Sciences*, **116**(34), 17115.

Stern, A. J., Wilton, P. R., and Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from dna sequence data. *PLoS Genetics*, **15**(9), 1–32.

Terhorst, J., Kamm, J. A., and Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, **49**(2), 303–309.

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

Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Roberts & Co. Publishers.

Wang, K., Mathieson, I., O'Connell, J., and Schiffels, S. (2020). Tracking human population structure through time from whole genome sequences. *PLoS Genetics*, **16**(3), 1–24.

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

Appendix

A PDE Solution Scheme

To numerically compute the requisite solutions to the PDEs (14) and (17), we implement the solution scheme outlined by Miroshnikov and Steinrücken (2017). 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 (14) here, but the same scheme can be used for the emission distribution (17). 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 $0 \le t \le \frac{t_{S-1}}{2}$, $v^a(\sigma) \cdot t \le x \le n \cdot t$, and $v^b(\sigma) \cdot t \le x \le n \cdot t$. The grid consists of slices in the x-y plane that are spaced across t according to a specified discretization $D := \left\{t_0 = 0, \ldots, t_K = \frac{t_{S-1}}{2}\right\}$ into K values. Each successive grid along t is defined by propagating along the direction of the characteristic $(x_0 + v^a(\sigma)\tau, y_0 + v^b(\sigma))$ 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 $x_0 = n \cdot t_i$ or $y_0 = n \cdot t_i$. 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:

$$
\begin{aligned}
F_\sigma\big(t_0 + \tau, x_0 + v^a(\sigma)\tau, y_0 + v^b(\sigma)\tau\big) &= e^{-H_k(\tau)}\Big(F_{\sigma 0} + \int_0^\tau g_\sigma(\alpha) e^{H_k(\alpha)} d\alpha\Big) \\
\text{where } g_\sigma(\tau) &:= \sum_{\sigma' \to \sigma} F_{\sigma'}\big(t_0 + \tau, x_0 + v^a(\sigma)\tau, y_0 + v^b(\sigma)\tau\big) Q^\rho_{\sigma',\sigma}(t)(t_0 + \tau), \\
H_\sigma(\tau) &:= \int_0^\tau q_\sigma(\alpha) d\alpha, \\
q_\sigma(\tau) &:= -Q^\rho_{\sigma,\sigma}(t_0 + \tau), \text{ and} \\
&\sigma' \to \sigma \text{ is the set of states } \sigma' \text{ that precede } \sigma.
\end{aligned}
$$

This scheme implementing the Methods of Characteristic propagates the boundary values along the characteristics. While the integral for $H_\sigma(\tau)$ is computed exactly (since we can analytically integrate $\lambda(t)$ for spline and piecewise constant representations) 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 (14), the values along the boundaries are the solution to another set of PDEs (15), 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 $F_\sigma(t, x = y = n \cdot t) = g^\rho_\sigma(t)$ 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.

We furthermore improve efficiency by imposing the following symmetries and constraints explicitly in our solution scheme, 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 (16) to compute the requisite probabilities for the transitions of our CHMM. Emission probabilities are computed using an analagous scheme for the PDE given in (17).

B Population Size Parameters

B.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 in two different ways.

Our default option is 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 also include an option to perform the inference using a cubic-spline representation of the population size history, similar to that offered by `SMC++` Terhorst et al. (2017). 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.

B.2 Initialization of Population Size History

We initialize the EM algorithm to infer the population size history such that $N(k) = \hat{N}$ is constant over time. 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 (Appendix C) and in computing regularizing coefficients (Appendix B.2).

B.3 Regularization of Inference Procedure

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 (Terhorst et al., 2017).

We enforce these regularization constraints into our method by modifying the objective function during the M-step. The parameter update during EM (from (18)) then becomes

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

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.

C 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 divide the continuum into equidistant intervals on a logarithmic or geometric scale (e.g. Li and Durbin, 2011). 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 Wakeley (2009). By numerically evaluating these distributions, we find the respective uniform partitions.

Since the computing time required for the forward-backward algorithm scales quadratically with the number of hidden states (matrix multiplication), even using 50 states may create a computational bottleneck. However, using the computational improvements detailed in Section D can compensate and make the algorithm tractable again by reducing the number of times matrix multiplication is performed.

D Improving Speed for Forward and backward Algorithm

Computing the entries of the full forward and backward table for each nucleotide site of the genome using the respective algorithm would be highly inefficient, and thus we implement two different ways to improve here. The *locus-skipping* efficiently integrates over large monomorphic regions without segregating sites, whereas the *meta-locus model* approximates the likelihood to reduce the number of "effective" loci for the CHMM.

D.1 Locus-Skipping

The *locus-skipping* method to improve efficiency was implemented as described in the supplement of Terhorst et al. (2017). 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 a 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, it retains all the relevant information necessary to compute the expectations required in the EM algorithm.

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. Tracts of missing data are straightforward to handle in this case as well, since can be treated as monomorphic tracts with modified emission probabilities, where each hidden state is equally likely to emit a missing nucleotide.

D.2 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 (Steinrücken et al., 2019, 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, 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 D.1 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).

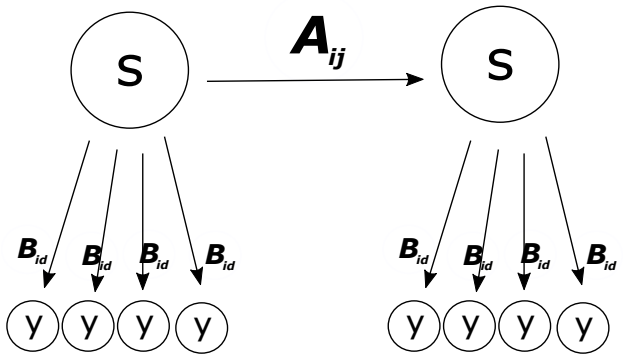


Figure D.1: Schematic for *meta-locus* model with $s = 4$. Each meta-locus represents 4 sites. Recombination does not happen within a meta-locus, but between at elevated rates. The emission probabilities $\mathbf{B}_{id}$, 4 for each meta-locus, are the same as in the single-step model.

This implementation offers the potential for a tremendous speedup of the forward-backward algorithm and this option should be used used when the E-step appears to be the computational bottleneck during inference, in particular when the number of discrete CHMM states is large.

E 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 method 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 to provide to our CHMM. Lastly, our method needs information on whether a certain allele at a segregating site is ancestral or derived, which can be specified by providing an ancestral sequence in addition to the population sample.

E.1 Missing Data

E.1.1 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.

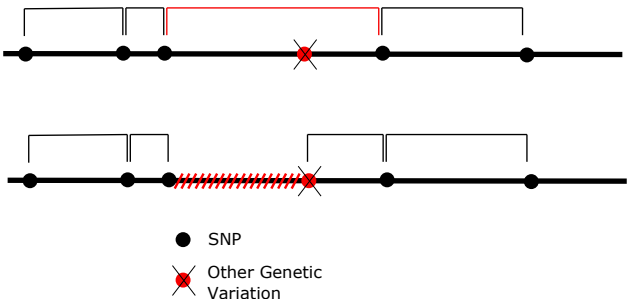


Figure E.1: This schematic shows the difference between just tagging a non bi-allelic segregating site as missing and also masking the tract upstream as missing. In the former case (top), the tract is effectively interpreted as an artificially long region without variation (marked red), which biases the inference. By masking the left region as missing (bottom) we avoid this systematic bias.

E.1.2 Tagging

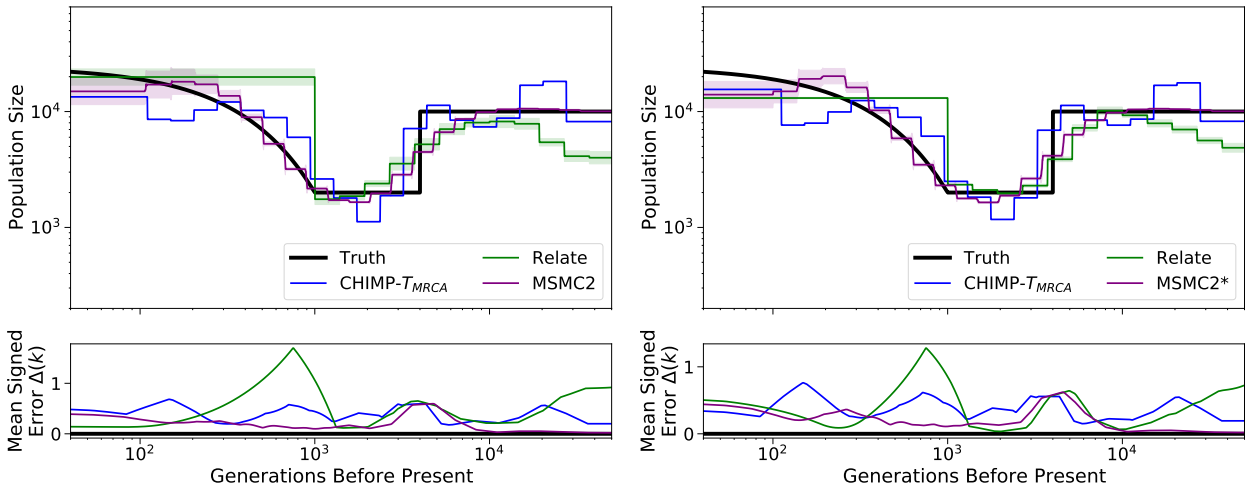
We note that genomic datasets can contain segregating sites that are not bi-allelic, for example, segregating sites with 3 alleles, or some forms of structural variation. We treat all segregating sites that are not bi-allelic as *missing* sites. This step introduces a potential bias into the CHMM because 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 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 E.1 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.

F Heuristics for Discretization of Population Size History

Figure F.1 shows the results when inferring the population size history in the bottleneck followed by exponential growth scenario (see Section 3.3), 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.

G Simulations with Recombination Map

Here, we explore the effects of varying recombination rates along the chromosome on demographic inference using the different mathods. We used `msprime` to simulate data under the *sawtooth* demography (explored in Section 3.3) using the HapMap II recombination rates on Chromosome 3 (The International HapMap Consortium, 2007). 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 in Figure 8. We provided the methods with as much information about the recombination rates as possible: we provided `Relate` with the



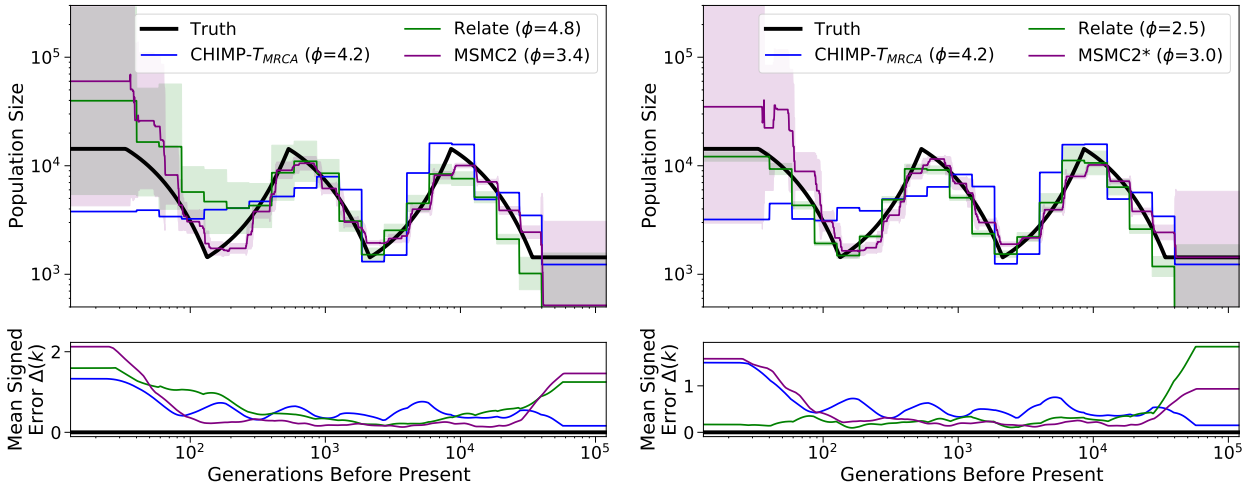
(a) Sample size 10.

(b) Sample size 200.

Figure F.1: 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.

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 G.2. Comparing these results to those in Figure 8, we note that `CHIMP`-T_{MRCA} and `MSMC2`, while operating under the assumption of a constant recombination rate, perform similarly well as in Figure 8, 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. 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.



(a) Sample size 10.

(b) Sample size 200.

Figure G.2: 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.