Bayesian Nonparametric Inference of Population Size Changes from Sequential Genealogies

Julia A. Palacios[1,2,3*], John Wakeley[1] and Sohini Ramachandran[2,3,*]
[1]Department of Organismic and Evolutionary Biology, Harvard University
[2]Department of Ecology and Evolutionary Biology, Brown University
[3]Center for Computational Molecular Biology, Brown University

Abstract

Sophisticated inferential tools coupled with the coalescent model have recently emerged for estimating past population sizes from genomic data. Accurate methods are available for data from a single locus or from independent loci. Recent methods that model recombination require small sample sizes, make constraining assumptions about population size changes, and do not report measures of uncertainty for estimates. Here, we develop a Gaussian process-based Bayesian nonparametric method coupled with a sequentially Markov coalescent model which allows accurate inference of population sizes over time from a set of genealogies. In contrast to current methods, our approach considers a broad class of recombination events, including those that do not change local genealogies. We show that our method outperforms recent likelihood-based methods that rely on discretization of the parameter space. We illustrate the application of our method to multiple demographic histories, including population bottlenecks and exponential growth. In simulation, our Bayesian approach produces point estimates four times more accurate than maximum likelihood estimation (based on the sum of absolute differences between the truth and the estimated values). Further, our method's credible intervals for population size as a function of time cover 90 percent of true values across multiple demographic scenarios, enabling formal hypothesis testing about population size differences over time. Using genealogies estimated with *ARGweaver*, we apply our method to European and Yoruban samples from the 1000 Genomes Project and confirm key known aspects of population size history over the past 150,000 years.

keywords. Markov Process, Genomics, Sequentially Markov Coalescent, Point Process, Gaussian Process.

1 Introduction

For a single non-recombining locus, neutral coalescent theory predicts the set of timed ancestral relationships among sampled individuals, known as a gene genealogy (Kingman 1982; Hudson 1983; Tajima 1983; Hudson 1990). In the coalescent model with variable population size, the rate at which two lineages coalesce, or have a common ancestor, is a function of the population size in the past. Here we denote the *population size trajectory* by $N(t)$, where t is time in the past, and use the term *local genealogy* to describe ancestral relationships at one non-recombining locus.

[*] corresponding authors: juliapalaciosroman@fas.harvard.edu, sramachandran@brown.edu

When analyzing multilocus sequences, a single local genealogy will not represent the full history of the sample. Instead, the set of ancestral relationships and recombination events among a sample of multilocus sequences can be represented by a graph, known as the ancestral recombination graph (ARG) which depicts the complex structure of neighboring local genealogies and results in a computationally expensive model for inferring $N(t)$ (Griffiths and Marjoram 1997; Wiuf and Hein 1999).

Recent studies have leveraged computationally simpler approximations for the coalescent with recombination—the sequentially Markov coalescent (SMC) (McVean and Cardin 2005) and its variant SMC$'$ (Marjoram and Wall 2006; Chen et al. 2009)—both of which model local genealogies as a continuous time Markov process along sequences (Figure 1). The difference between the SMC and SMC$'$ is that the SMC models only the class of recombination events that alter local genealogies of the sample. In general, the SMC$'$ is a better approximation to the ARG than the SMC (Chen et al. 2009; Wilton et al. 2015). Because of these features, in this work we rely on the SMC$'$ to model local genealogies with recombination.

Under the coalescent and the sequentially Markov coalescent (SMC and SMC$'$) models, population size trajectories and sequence data are separated by two stochastic processes: *i*) *a state process* which describes the relationship between the population size trajectory and the set of local genealogies, and *ii*) *an observation process* which describes how the hidden local genealogies are observed through patterns of nucleotide diversity in the sequence data. The observation process includes mutation and genotyping error while the state process models coalescence. Sequence data are then used to make inferences of population size trajectories. In this paper, we restrict attention to the state process of local genealogies and show how inferences of population size trajectories can be made from them. We solve a number of key modeling and inference problems, and thus provide a basis for developing efficient algorithms to infer population parameters from sequence data directly.

Whole-genome inference of population size trajectories has been hampered by the enormous size of the state space of local genealogies when the sample size is large. The pioneering, pairwise sequentially Markov Coalescent (PSMC) method of Li and Durbin (2011) employed the SMC to make inferences from a sample of size two ($n = 2$). In this method, time is discretized and the population size trajectory is piece-wise constant, allowing pairwise genealogies also to be discretized. Subsequent methods for samples larger than two similarly rely on the discretization of time and genealogies. The natural extension of the PSMC to $n > 2$ is the multiple sequentially Markovian coalescent (MSMC) (Schiffels and Durbin 2014). However, the MSMC models only the most recent coalescent event of the sample, and hence its estimation of population sizes is limited to very recent times. Other recent methods propose efficient ways of exploring the state space of hidden genealogies for $n > 2$ (Sheehan et al. 2013; Rasmussen et al. 2014), yet also rely on discretizing the state space of local genealogies and assume a piece-wise constant trajectory of population sizes. We show that the *a priori* specification of change points for the piece-wise population size trajectory required by current approaches is problematic because estimates of $N(t)$ are sensitive to this specification. Moreover, current methods do not generate interval estimates for $N(t)$.

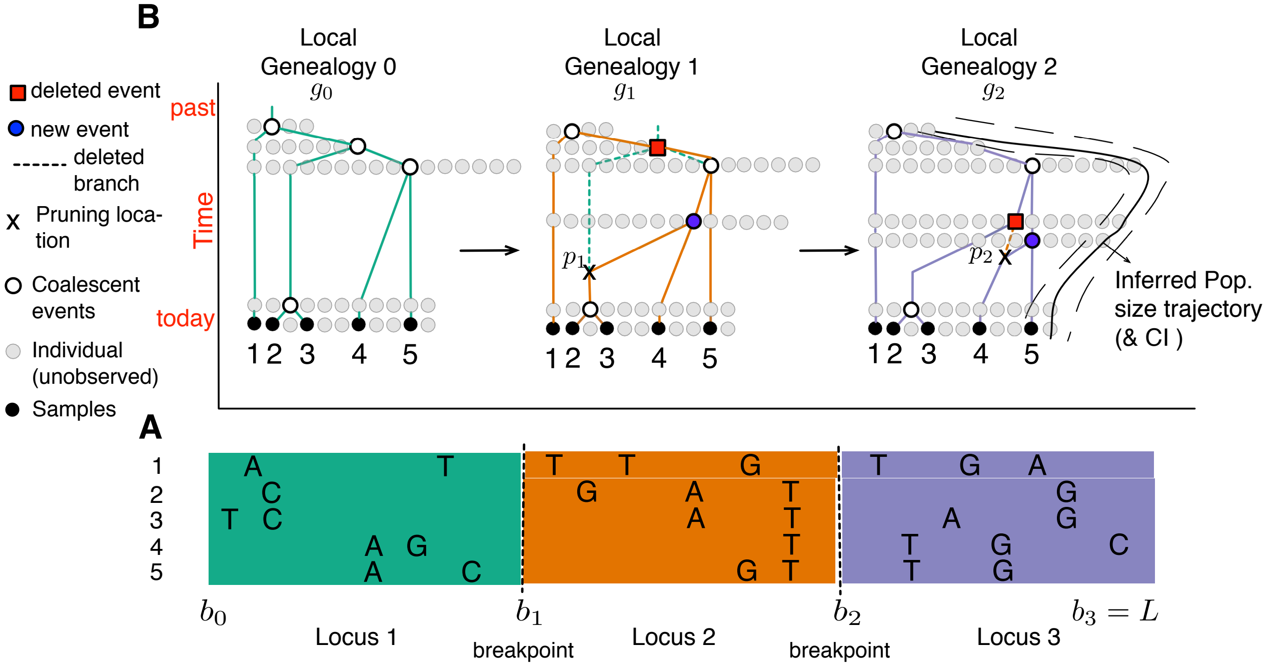


Figure 1: **SMC′ model for inferring population size trajectories**. Drawn after Rasmussen et al. (2014) to highlight notation specific to our study. **A.** Observed sequence data in a segment of length L from five individuals; three loci are shown delimited by recombination breakpoints b_1 and b_2. Only the derived mutations at polymorphic sites are shown. **B.** Corresponding local genealogies g_i for each locus i. The five sampled individuals are depicted as black filled circles. Local genealogies have a Markovian degree 1 dependency. Each inter-coalescent time (the time interval between coalescent events denoted as empty circles) provides information about past population size (number of gray filled circles at a given time point). Moving from left to right after recombination breakpoint b_1, the pruning location p_1 is selected from genealogy g_0 and the pruned branch is regrafted back on the genealogy (blue filled circle). The coalescent event of g_0 depicted as a red filled circle in g_1 is deleted. This creates the next genealogy g_1. The process continues until L. At L, the population size trajectory $N(t)$ (depicted as a black curve superimposed on g_2) can be inferred.

Gaussian Process-based Bayesian inference of population size trajectories has proven to be a powerful and flexible nonparametric approach when applied to a single local genealogy (Palacios and Minin 2013; Lan et al. 2015). The two main advantages of the GP-based approach are: (*i*) it does not require a specific functional form of the population size trajectory (such as constant or exponential growth) and (*ii*) it does not require an arbitrary specification of change points in a piece-wise constant or linear framework.

In this paper, we show the downstream effects of discretizing time, assuming a piecewise constant trajectory, and reporting only point estimates for past population sizes. We overcome previous limitations by introducing a Bayesian nonparametric approach with a Gaussian Process (GP) to model the population size trajectory as a continuous function of time. More specifically, we model the logarithm of the population size trajectory *a priori* as a Gaussian process (the log ensures our estimates are positive). As mentioned above, we assume that local gene genealogies are known. For our Bayesian model, we develop a Markov Chain Monte Carlo (MCMC) method to sample from the posterior distribution of population sizes over time. Our MCMC algorithm uses the recently developed algorithm Split Hamiltonian Monte Carlo (splitHMC) (Shahbaba et al. 2014; Lan et al. 2015). splitHMC updates all model parameters jointly and it can be extended to a full inferential framework that is directly applicable to sequence data. In order to compare our Bayesian GP-based estimation of population size trajectories with a piece-wise constant maximum likelihood-based estimation (e.g. Li and Durbin 2011; Sheehan et al. 2013; Schiffels and Durbin 2014), we implemented the Expectation-maximization (EM) algorithm within our framework and computed the observed Fisher information to obtain confidence intervals of the maximum likelihood estimates.

Lastly, we address a key problem for inference of population size trajectories under sequentially Markov coalescent models is the efficient computation of transition densities needed in the calculation of likelihoods. Here, we express the transition densities of local genealogies in terms of local ranked tree shapes (Tajima 1983) and coalescent times, and show that these quantities are statistically sufficient for inferring population size trajectories either from sequence data directly or from the set of local genealogies. The use of ranked tree shapes allows us to exploit the state process of local genealogies efficiently since the space of ranked tree shapes has a smaller cardinality than the space of labeled topologies (Sainudiin et al. 2014).

2 Methods: SMC$'$ Calculations

Following notation similar to Rasmussen et al. (2014) (Table 1) a realization of the embedded SMC$'$ chain consists of a set of m local genealogies $(g_0, g_1, \ldots, g_{m-1})$, $m-1$ recombination breakpoints at chromosomal locations $(b_1, b_2, \ldots, b_{m-1})$, and $m-1$ pruning locations $(p_1, p_2, \ldots, p_{m-1})$, where $p_i = (u_i, w_i)$ indicates the time of the recombination event u_i and the branch w_i where recombination happened in genealogy g_{i-1} (Figure 1). Genealogy g_0 corresponds to the genealogy of n sequences that contains the set of timed ancestral relationships among the n individuals for the chromosomal segment $(0, b_1]$. Genealogy g_i corresponds to the genealogy of the same n sequences for the chromosomal segment $(b_i, b_{i+1}]$ for $i = 1, 2, \ldots, m-2$. Finally, t_j^i denotes the time when two of j lineages coalesce in genealogy g_i, measured in units of generations before present.

Using capital letters to denote random variables, the evolution of the SMC′ process along chromosomal segments is governed by a point process $B = \{B_i\}_{i\in\mathbb{N}}$ that represents the random locations of recombination breakpoints. We use $S_i = B_i - B_{i-1}$, for $i = 1, 2, \ldots, m$, to denote the segment lengths for each local genealogy, with $S_0 = B_0 = 0$. Let $G = \{G_i\}_{i\in\mathbb{N}}$ be the chain which records the local genealogies, and let $P = (U, W) = \{(U_i, W_i)\}_{i\in\mathbb{N}}$ represent the chain which records the pruning locations (time and branch) on G. The sequence $(G_i, P_i = \{U_i, W_i\}, B_i)$ has the following conditional independence relation:

$$\Pr[G_i = g_i, U_i \le u_i, W_i = w_i, S_i \le s \mid (g_0, b_0), (g_1, u_1, w_1, b_1), \ldots, (g_{i-1}, u_{i-1}, w_{i-1}, b_{i-1})]$$

$$= \Pr[S_i \le s_i \mid g_{i-1}] \tag{1}$$

$$\times \Pr[U_i \le u_i, W_i = w_i \mid g_{i-1}] \tag{2}$$

$$\times \Pr[G_i = g_i \mid U_i \le u_i, W_i = w_i, g_{i-1}] \tag{3}$$

Given a chain of local genealogies, pruning locations and recombination breakpoints, the joint transition probability to a new genealogy, pruning location and locus length can be expressed as the product of the locus length probability conditioned on the current genealogy (Expression 1, above), the pruning location probability conditioned on the current genealogy (Expression 2, above) and, the transition probability of the new genealogy conditioned on the current genealogy and pruning location (Expression 3, above).

2.1 Complete data transition densities

Consider the chain of local genealogies $\mathbf{g} = (g_0, g_1, \ldots, g_{m-1})$ with recombination breakpoints at $\mathbf{b} = (0, b_1, \ldots, b_{m-1})$. According to the SMC′ process, the first local genealogy g_0 follows the standard coalescent density:

$$\Pr[G_0 = g_0 \mid N(t)] = \prod_{j=2}^{n} \frac{1}{N(t_j^0)} \exp\left\{ - \int_{t_{j+1}^0}^{t_j^0} \frac{A^0(t)(A^0(t)-1)dt}{2N(t)} \right\}, \tag{4}$$

where $t_{n+1}^0 = 0$ and $t_n^0 < \ldots < t_2^0$ are the set of coalescent times in local genealogy g_0. The piece-wise constant function $A^i(t)$ denotes the number of ancestral lineages present at time t in genealogy g_i, that is

$$A^i(t) = \sum_{j=1}^{n} j 1_{t\in(t_{j+1}^i, t_j^i)},$$

with $t_1^i = \infty$.

Table 1: Notation for the SMC$'$ model used in this work.

	Symbol	Description
Parameters:	ρ	Recombination rate per site per generation
	$N(t)$	Effective population size trajectory with time measured in units of N_0 generations
	τ	Hyperparameter that controls the smoothness of the log-Gaussian process prior on $N(t)$
Notation specific to SMC$'$ chain:	L	Length of observed sequences
	b_i	Chromosomal location of the ith recombination breakpoint
	m	Number of local genealogies corresponding to $m-1$ recombination events
	$s_{i+1} = b_{i+1} - b_i$	Segment length for local genealogy i
	g_i	Local genealogy for the segment (b_{i-1}, b_i)
Notation specific to local genealogy:	n	Sample size, or number of sequences
	l_i	Total tree length of local genealogy g_i
	$A^i(t)$	Piece-wise constant function of the number of ancestral lineages at time t in local genealogy g_i
	t^i_j	Coalescent time in genealogy g_i when two of j lineages coalesce. $A^i(t^i_j-) = j$; $A^i(t^i_j+) = j-1$
	$\mathbf{t}^i = (t^i_n, t^i_{n-1}, \ldots, t^i_2)$	Vector of coalescent times of genealogy g_i
	$p_i = (u_i, w_i)$	Pruning location along local genealogy g_i
	u_i	Time when the recombination event happened along the height of the genealogy g_i
	w_i	Lineage on genealogy g_{i-1} where the recombination event happened
	w'_i	New lineage added on genealogy g_i where the recombination event happened
	t^i_{new}	Coalescent time in genealogy g_i when the lineage w_i coalesces.
	t^i_{del}	Coalescent time in genealogy g_{i-1} that no longer exists in genealogy g_i
	c_i	Lineage on genealogy g_i that coalesces with lineage w'_i
	$F^i_{j,k}$	Number of free lineages in local genealogy g_i that do not coalesce in the time interval (t^i_{j+1}, t^i_k)
	$I^i(t)$	Piece-wise constant function that takes values in $\{0,1,2\}$ indicating the number of ancestral lineages at time t in genealogy g_i where the pruning event would produce a visible transition to g_{i+1}
Discretization:	d	Number of change points at which $N(t)$ is estimated
	$\mathbf{x} = (x_1, \ldots, x_d)$	Times at which $N(t)$ is estimated

Given a current local genealogy g_{i-1}, the distribution of the length $S_i = B_i - b_{i-1}$ of the current locus depends on the current state of the SMC′ chain through the local genealogy's total tree length l_{i-1} (the sum of all branch lengths in g_{i-1}) and the recombination rate per site per generation ρ.

$$f(s_i \mid g_{i-1}, \rho) = \rho l_{i-1} \exp\{-\rho l_{i-1} s_i\}. \tag{5}$$

At recombination breakpoint b_i, a new local genealogy g_i is generated (Figure 1). This new local genealogy g_i depends on the previous local genealogy g_{i-1} and the population size trajectory $N(t)$. To generate g_i we first randomly choose a pruning location p_i (consisting of a pruning time u_i and a lineage w_i) uniformly along g_{i-1}. At pruning location p_i, we add a new lineage w'_i and coalesce it further in the past at time t^i_{new} with some lineage, c_i (Figure 2). We then delete the w_i lineage's segment from u_i to t^i_{del} (the coalescent time of lineage w_i). The transition density to a new genealogy at recombination breakpoint b_i is then

$$\begin{aligned}\Pr[p_i = (u_i, w_i), t^i_{new}, c_i \mid g_{i-1}, N(t)] &= \Pr[p_i = (u_i, w_i) \mid g_{i-1}] \Pr[t^i_{new}, c_i \mid u_i, g_{i-1}, N(t)] \\ &= \left(\frac{1}{l_{i-1}}\right) \frac{1}{N(t^i_{new})} \exp\left\{-\int_{u_i}^{t^i_{new}} \frac{A^{i-1}(t)dt}{N(t)}\right\},\end{aligned} \tag{6}$$

where l_{i-1} denotes the total tree length of g_{i-1}.

This generative process of local genealogies can result in the two types of transitions depicted in Figure 2. A *visible transition* results in a genealogy g_i which is different from g_{i-1} (Figure 2A), while an *invisible transition* makes g_i identical to g_{i-1} (Figure 2B).

An invisible transition $g_i = g_{i-1}$, occurs when $c_i = w_i$. Given the pruning location $p_i = (u_i, w_i)$, a transition to an invisible event occurs when $T^i_{new} \in (u_i, t^i_{del})$ and C_i, the random variable indicating the lineage that coalesces with lineage w'_i, takes the value w_i. The probability of an invisible transition is given by

$$\begin{aligned}\Pr[G_i = g_{i-1} \mid p_i = (u_i, w_i), g_{i-1}, N(t)] &= \Pr[u_i \le T^i_{new} \le t^i_{del}, C_i = w_i \mid p_i = (u_i, w_i), g_{i-1}, N(t)] \\ &= \int_{u_i}^{t^i_{del}} \frac{1}{N(t)} \exp\left\{-\int_{u_i}^{t} \frac{A^{i-1}(u)du}{N(u)}\right\} dt.\end{aligned}$$

Thus, the joint transition probability to an invisible event with pruning location (u_i, w_i), given g_{i-1} is:

$$\Pr[G_i = g_{i-1}, p_i = (u_i, w_i) \mid g_{i-1}, N(t)] = \frac{1}{l_{i-1}} \Pr[G_i = g_{i-1} \mid p_i = (u_i, w_i), g_{i-1}, N(t)].$$

2.2 Transition densities averaged over unknown pruning locations

Even though we will assume that local genealogies are known, in order to anticipate later applications to sequence data we do not wish to make the same assumption about pruning locations. Thus, we average over pruning locations to obtain marginal transition densities between genealogies, for both visible and invisible transitions.

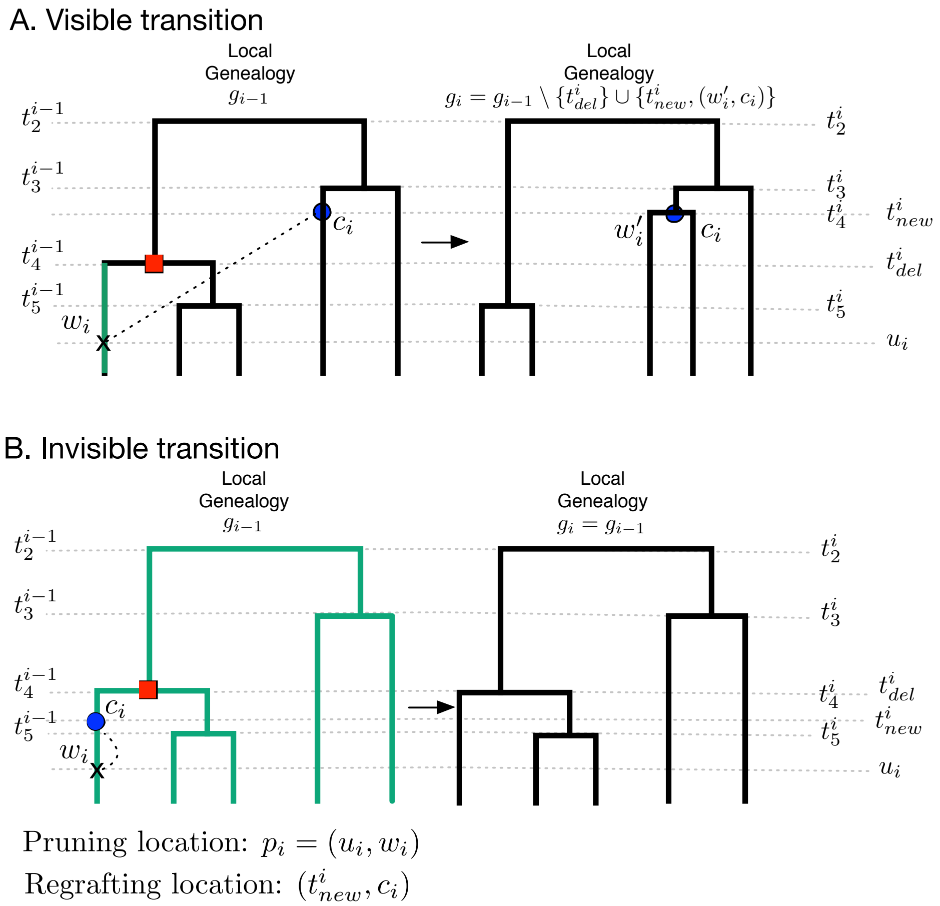


Figure 2: **Schematic representation of SMC′ transitions given a recombination breakpoint at location b_i (indicated as an arrow in each panel). A: Visible transition.** We uniformly sample the pruning location p_i from g_{i-1} at time u_i along some branch w_i, we add a new branch w'_i at u_i and re-graft it (dashed black line). The new branch w'_i coalesces with some branch c_i at time t^i_{new}. We then delete branch w_i and the coalescent time t^i_{del} to generate genealogy g_i. Any pruning time along the branch w_i (shown in green) would have produced the same visible transition from g_{i-1} to g_i. **B: Invisible transition.** We uniformly sample the pruning location $p_i = (u_i, w_i)$, add a new branch w'_i at u_i and re-graft it. The new branch w'_i coalesces with itself (dashed black line); that is, $C_i = w_i$, and then the segment (u_i, t^i_{del}) of w_i is deleted. If $C_i = w_i$, any pruning location along the green branches would have produced the same invisible transition.

To compute the marginal visible transition density to a new genealogy $g_i = \{g_{i-1} \setminus \{t^i_{del}\} \cup \{t^i_{new}, (w'_i, c_i)\}\}$, we need to average over all possible pruning locations $p_i = (u_i, w_i)$ along g_{i-1}. By comparing the two genealogies g_{i-1} and g_i in Figure 2A, we know that p_i corresponds to the lineage w_i some time along $(0, t^{i-1}_4)$, or equivalently, along $(0, t^i_{del})$. In general, comparison of g_{i-1} and g_i may not provide complete information to identify the lineage that was pruned. When the children of the node corresponding to t_{del} and the children of the node corresponding to t_{new} are the same, pruning different branches can lead to the same transition. We enumerate all cases of incomplete information for visible transitions in Supporting Information Figure S1.

We introduce a function $I^{i-1}(t)$, equal to the number of possible lineages at time t where the pruning location along g_{i-1} would produce a visible transition to g_i. $I^{i-1}(t)$ is a piece-wise constant function that takes the values in $\{0, 1, 2\}$ depending on whether the pruning location p_i can happen in $0, 1$ or 2 branches at time t. In the example in Figure 2A,

$$I^{i-1}(t) = \begin{cases} 1, & \text{if } t \in (0, t^{i-1}_4), \\ 0, & \text{if } t \in (t^{i-1}_4, \infty). \end{cases} \tag{7}$$

For a general $I^{i-1}(t)$ piece-wise constant function that indicates the number of possible pruning branches at time t, the marginal visible transition density to a new genealogy is

$$\begin{aligned} \Pr[G_i = g_i \mid g_{i-1}, N(t)] &= \frac{1}{l_{i-1}} \int_0^\infty I^{i-1}(u) \Pr[t^i_{new}, c_i \mid u, w_i] du \\ &= \frac{1}{l_{i-1}} \int_0^\infty I^{i-1}(u) \frac{1}{N(t^i_{new})} \exp\left\{ - \int_u^{t^i_{new}} \frac{A^{i-1}(t)dt}{N(t)} \right\} du. \end{aligned} \tag{8}$$

Turning now to the computation of marginal transition probabilities for invisible events, we need to average over all possible pruning locations p_i. Consider the example in Figure 2B and choosing a pruning time (u_i) along g_{i-1}. In order to have an invisible transition, the coalescing branch C_i must be the same pruning branch W_i. In Figure 2B the new coalescent time T^i_{new} can happen along five lineages in the interval $(0, t^{i-1}_5)$, three lineages in the interval (t^{i-1}_5, t^{i-1}_4), and two lineages in the interval (t^{i-1}_4, t^{i-1}_3). To generalize this calculation, we introduce the quantity $F^i_{j,k}$ with $(n+1) \geq j \geq k \geq 2$ which denotes the number of lineages in g_i that are *free* (do not coalesce), in the time segment (t^i_{j+1}, t^i_k). The time interval (t^i_{j+1}, t^i_k) includes the interval of pruning (t^i_{j+1}, t^i_j) up to the interval of self-coalescence (t^i_{k+1}, t^i_k). Thus, if the pruning time happens at time $U_i \in (t^i_j, t^i_{j-1})$, an invisible transition with new coalescent time $T^i_{new} \in (t^i_{k+1}, t^i_k)$ can happen along $F^i_{j,k}$ free lineages.

In Figure 2B, u_i happened in the time interval $(0, t^{i-1}_5)$. If the new coalescent time T^i_{new} happens in the interval (u_i, t^{i-1}_5) along the same (unknown) pruning branch, then this invisible transition has probability

$$\Pr[G_i = g_{i-1}, T^i_{new} \in (t^{i-1}_6, t^{i-1}_5) \mid u_i, g_{i-1}, N(t)] = F^{i-1}_{5,5} \int_{u_i}^{t^{i-1}_5} \frac{1}{N(t)} \exp\left\{ - \int_{u_i}^{t} \frac{A^{i-1}(u)du}{N(u)} \right\} dt,$$

with $F_{5,5} = 5$.

Now consider the same example of Figure 2B but with an unknown pruning time u_i. The joint event where recombination occurs at pruning time $U_i \in (t_6^{i-1}, t_5^{i-1})$ and coalescent time T_{new}^i occurs in the interval (t_6^{i-1}, t_5^{i-1}) and this results in an invisible transition has probability:

$$\Pr[G_i = g_{i-1}, U_i \in (t_6^{i-1}, t_5^{i-1}), T_{new}^i \in (t_6^{i-1}, t_5^{i-1}) \mid g_{i-1}, N(t)]$$

$$= \frac{F_{5,5}^{i-1} \int_{t_6^{i-1}}^{t_5^{i-1}} \int_{u_i}^{t_5^{i-1}} \frac{1}{N(t)} \exp\left\{- \int_{u_i}^{t} \frac{A^{i-1}(u)du}{N(u)}\right\} dt du_i}{l_{i-1}} \tag{9}$$

$$= \frac{F_{5,5}^{i-1} P_{5,5}^{i-1}}{l_{i-1}}, \tag{10}$$

where $P_{5,5}^{i-1}$ denotes the double integral expression in Equation 9 for ease of notation.

An invisible transition would also result if $U_i \in (t_6^{i-1}, t_5^{i-1})$ and $T_{new}^i \in (t_5^{i-1}, t_4^{i-1})$ along the same (unknown) pruning branch; according to Figure 2B, this can happen along three lineages, so $F_{5,4}^{i-1} = 3$ and this event has probability:

$$\begin{aligned}
&\Pr[G_i = g_{i-1}, U_i \in (t_6^{i-1}, t_5^{i-1}), T_{new}^i \in (t_5^{i-1}, t_4^{i-1}) \mid g_{i-1}, N(t)] \\
&= \frac{F_{5,4}^{i-1} \int_{t_6^{i-1}}^{t_5^{i-1}} \exp\left\{- \int_{u_i}^{t_5^{i-1}} \frac{A^{i-1}(u)du}{N(u)}\right\}}{l_{i-1}} \int_{t_5^{i-1}}^{t_4^{i-1}} \frac{1}{N(t)} \exp\left\{- \int_{t_5^{i-1}}^{t} \frac{A^{i-1}(u)du}{N(u)}\right\} dt du_i \\
&= \frac{F_{5,4}^{i-1} P_{5,4}^{i-1}}{l_{i-1}}.
\end{aligned}$$

If we continue considering the cases where $U_i \in (t_6^{i-1}, t_5^{i-1})$ and $T_{new}^i \in (t_4^{i-1}, t_3^{i-1})$ or $T_{new}^i \in (t_3^{i-1}, t_2^{i-1})$, we have $F_{5,3}^{i-1} = 2$ and $F_{5,2}^{i-1} = 0$. Then, the joint probability of an invisible event and $U_i \in (t_6^{i-1}, t_5^{i-1})$ is

$$\Pr[G_i = g_{i-1}, U_i \in (t_6^i, t_5^i) \mid g_{i-1}, N(t)] = \frac{\sum_{k=2}^{6} F_{j,k}^{i-1} P_{j,k}^{i-1}}{l_{i-1}},$$

For the cases when $U_i \in (t_{j+1}^{i-1}, t_j^{i-1})$ and the new coalescent time T_{new}^i falls in another coalescent interval $(t_{k+1}^{i-1}, t_k^{i-1})$, we need to compute the following:

- The joint probability of $U_i \in (t_{j+1}^{i-1}, t_j^{i-1})$ and no coalescence in the interval (u_i, t_j^{i-1}):

$$\frac{1}{l_{i-1}} Q_j^{i-1} = \frac{1}{l_{i-1}} \int_{t_{j+1}^{i-1}}^{t_j^{i-1}} \exp\left\{- \int_{u_i}^{t_j^{i-1}} \frac{C^{i-1}(u)du}{N(u)}\right\} du_i,$$

- The probability of no coalescence in any of the intermediate coalescent intervals $(t_{l+1}^{i-1}, t_l^{i-1})$:

$$q_l^{i-1} = \exp\left\{- \int_{t_{l+1}^{i-1}}^{t_l^{i-1}} \frac{C^{i-1}(u)du}{N(u)}\right\},$$

and

- The probability of coalescing at $T^i_{new} \in (t^{i-1}_{k+1}, t^{i-1}_k)$:

$$1 - q^{i-1}_k.$$

Then,

$$\frac{1}{l_{i-1}} P^{i-1}_{j,k} = \frac{1}{l_{i-1}} Q^{i-1}_j q^{i-1}_{j-1} q^{i-1}_{j-2} \dots q^{i-1}_{k+1} (1 - q^{i-1}_k)$$

represents the probability that the pruning location is w_i at time $U_i \in (t^{i-1}_{j+1}, t^{i-1}_j)$ and the new lineage w'_i coalesces at time $T^i_{new} \in (t^{i-1}_{k+1}, t^{i-1}_k)$ with lineage $c_i = w_i$. Overall, the marginal transition probability to an invisible event is:

$$\begin{aligned}
\Pr[G_i = g_{i-1} \mid g_{i-1}, N(t)] &= \int_0^{t^{i-1}_2} \Pr[G_i = g_{i-1}, u_i \mid g_{i-1}, N(t)] du_i \\
&= \sum_{j=2}^{n} \Pr[G_i = g_{i-1}, U_i \in (t^{i-1}_{j+1}, t^{i-1}_j) \mid g_{i-1}, N(t)] \\
&= \frac{1}{l_{i-1}} \sum_{j=2}^{n} \sum_{k=2}^{j} F^{i-1}_{j,k} P^{i-1}_{j,k}. \qquad (11)
\end{aligned}$$

2.3 The likelihood of the embedded SMC′ chain

Instead of having a complete realization of the embedded SMC′ chain of m local genealogies $g_0, \dots, g_{m-1}$ and pruning locations $p_1, \dots, p_{m-1}$ at recombination breakpoints $b_1, \dots, b_{m-1}$, we assume that our data (unless otherwise noted) consist only of m local genealogies at recombination breakpoints from a chromosomal segment of length L. Note that our observed data are not sequence data. More specifically, our observed data are

$$\mathbf{Y} = \{(g_0, 0), (g_1, b_1) \dots, (g_{m-1}, b_{m-1}), s_m = L - b_{m-1}\}. \qquad (12)$$

Then, the observed data likelihood is

$$\mathcal{L}_{obs}(\mathbf{Y}; N(t), \rho) = \Pr[g_0 \mid N(t)] \left[\prod_{i=0}^{m-2} f[s_{i+1} \mid g_i, \rho] \Pr[g_{i+1} \mid g_i, N(t)] \right] h(L - b_{m-1} \mid g_{m-1}, \rho),$$

$$= \overbrace{\Pr[g_0 \mid N(t)] \left[\prod_{i=0}^{m-2} \Pr[g_{i+1} \mid g_i, N(t)] \right]}^{\text{factors that depend on } N(t)} \overbrace{h(L - b_{m-1} \mid g_{m-1}, \rho) \left[\prod_{i=0}^{m-2} f[s_{i+1} \mid g_i, \rho] \right]}^{\text{factors that depend on } \rho} \qquad (13)$$

where $h(L - b_{m-1} \mid g_{m-1}, \rho)$ is the survival function in state g_{m-1}. Equation 13 is factored into terms that depend on $N(t)$ alone and ones that depend on ρ alone. The terms that depend on ρ, given by Equation 5, depend on the data only through total tree lengths $l_0, \dots, l_{m-1}$ and locus lengths $s_1, \dots, s_{m-1}, L - b_{m-1}$. By the factorization theorem for sufficient statistics, local tree lengths $l_0, \dots, l_{m-1}$ and locus lengths $s_1, \dots, s_{m-1}, L - b_{m-1}$ are sufficient for inferring ρ.

3 Methods: Inference

Current coalescent-based methods that infer a population size trajectory $N(t)$ from whole-genome data assume $N(t)$ is a piece-wise constant function with change points $x_1 = 0 < x_2 < \ldots < x_d$ (Li and Durbin 2011; Sheehan et al. 2013; Rasmussen et al. 2014; Schiffels and Durbin 2014). That is

$$N(t) = \sum_{i=1}^{d} N_i 1_{t \in (x_{i-1}, x_i]}. \tag{14}$$

Equation 14 presents two challenges. The first challenge lies in the specification of the change points. The narrower an interval is, the higher the probability that we do not observe coalescent times in that interval. The fewer observed coalescent times in an interval, the greater the uncertainty of the estimate $\widehat{N}_i$ (if the estimate even exists). The second challenge lies in the specification of the time window $(0, x_d)$: if x_d is set too far in the past, we might not have enough data to accurately estimate $N(t)$ for $x_d \leq t < \infty$.

In order to solve the first challenge, Rasmussen et al. (2014) and Li and Durbin (2011) distribute the d change points evenly on a logarithmic scale:

$$x_j = \frac{1}{\kappa}\left\{\exp\left[\frac{j}{d}\log(1+\kappa x_d)\right] - 1\right\}. \tag{15}$$

where κ is specified by the user. Schiffels and Durbin (2014) propose discretizing time according to the quantiles of the exponential distribution.

$$x_j = \frac{-1}{\lambda}\log\left[1 - \frac{j}{d}\right], \tag{16}$$

where λ is the rate of an exponential distribution. Schiffels and Durbin (2014) model the time to the most recent coalescent event and set $\lambda = \binom{n}{2}$. However, Equation 16 is not directly applicable here because we use all coalescent events for inference.

In the following sections, we first present our Bayesian nonparametric method, then develop a maximum likelihood method under a piece-wise constant trajectory so we can directly compare an EM-based method (Li and Durbin 2011; Sheehan et al. 2013) to our Bayesian nonparametric method. The R code for all simulation studies and real data analysis conducted in this paper are publicly available at `http://ramachandran-data.brown.edu/datarepo/`.

3.1 Gaussian-Process-based Bayesian Nonparametric Estimation of $N(t)$

For our Bayesian methodology, we assume the following log-Gaussian Process prior on the population size trajectory, $N(t)$:

$$N(t) = \exp[f(t)],\ f(t) \sim \mathcal{GP}(\mathbf{0}, \mathbf{C}(\tau)), \tag{17}$$

where $\mathcal{GP}(\mathbf{0}, \mathbf{C}(\tau))$ denotes a Gaussian process with mean function $\mathbf{0}$ and inverse covariance function $\mathbf{C}^{-1}(\tau) = \tau\mathbf{C}^{-1}$ with precision parameter τ. For computational convenience, we use Brownian motion as our prior for $f(t)$ since its inverse covariance matrix is sparse. We place a Gamma prior on the precision parameter τ,

$$\tau \sim \Gamma(\alpha, \beta).$$

Assuming that recombination rate ρ is known, the posterior distribution of model parameters (Figure 3) is then

$$\Pr[N(t), \tau \mid g_0, \dots, g_{m-1}] \propto \Pr[g_0 \mid N(t)] \left\{ \prod_{i=0}^{m-2} \Pr[g_{i+1} \mid g_i, N(t)] \right\} \Pr[N(t) \mid \tau] \Pr(\tau). \tag{18}$$

The first two factors on the right side of Equation 18, detailed in Equations 8 and 11 involve integration over $N(t)$, an infinite dimensional random function (Equation 17). We approximate the integral

$$\int_a^b \frac{dt}{N(t)} = \int_a^b \exp[-f(t)]dt,$$

by the Riemann sum over a partition of the integration interval. That is,

$$\int_a^b \exp[-f(t)]dt \approx \sum_{j=i}^{k} \exp[-f_j^*]\Delta_j, \tag{19}$$

for $x_i < a < x_{i+1} < \dots < x_{k-1} < b < x_k$, $\Delta_i = x_{i+1} - a$, $\Delta_k = b - x_{k-1}$ and $\Delta_j = x_{j+1} - x_j$ for $i < j < k$. f_j^* is a representative value of $f(t)$ in the interval (x_j, x_{j+1}); in our implementation, we set $f_j^* = f(x_j^*)$ with $x_j^* = (x_j + x_{j+1})/2$. This way, we discretize our time window in d evenly spaced segments $x_1 = 0 < x_2 < \dots < x_d$, with $x_d = \max(t_1^0, \dots, t_1^{m-1})$, the maximum time to the most common ancestor observed in the sequence of local genealogies, and approximate $N(t)$ by a piece-wise linear function evaluated at $(x_1^*, x_2^*, \dots, x_d^*)$.

We condition on the set of m local genealogies $g_0, \dots, g_{m-1}$ to generate posterior samples for the vector $\mathbf{f}^* = [\log N(x_1^*), \dots, \log N(x_d^*)]$ and τ and use these posterior samples to infer $N(t)$ at $t \in (x_1^*, \dots, x_d^*)$, where $x_i^* = (x_i + x_{i+1})/2$. Updating $N(t)$ and τ separately is not recommended because of their strong dependency (Lan et al. 2015). Therefore, we update $(N(t), \tau)$ jointly in an MCMC sampling algorithm using Split Hamiltonian Monte Carlo (Shahbaba et al. 2014; Lan et al. 2015). Split Hamiltonian Monte Carlo relies on our ability to calculate the log-likelihood of the observed data and the gradient vector of the log-likelihood (i.e., the score function). The log-likelihood of the observed data is approximated via sums of the form in Equation 19. We approximate the score function $\nabla\mathcal{L}_{obs}(\mathbf{Y}; \mathbf{f}^*)$ with respect to $\mathbf{f}^*$ by applying Fisher's identity:

$$\nabla\mathcal{L}_{obs}(\mathbf{Y}; \mathbf{f}^*) = E_{\mathbf{f}^*}[\nabla\mathcal{L}_c(\mathbf{Y}_c; \mathbf{f}^*) \mid \mathbf{Y}],$$

where, at each iteration in the MCMC, expectation is calculated using the current value of $\mathbf{f}^*$. We show the details of this calculation in the Appendix.

Alternatively, one can update $N(t)$ in the MCMC algorithm using Elliptical Slice Sampler (Murray et al. 2010) with a fixed value of τ (perhaps estimated from previous studies or from a preliminary run from the Split Hamiltonian Monte Carlo algorithm). The advantage of using Elliptical Slice Sampler over the Split Hamiltonian Monte Carlo is purely computational since Elliptical Slice Sampler does not require calculation of the score function.

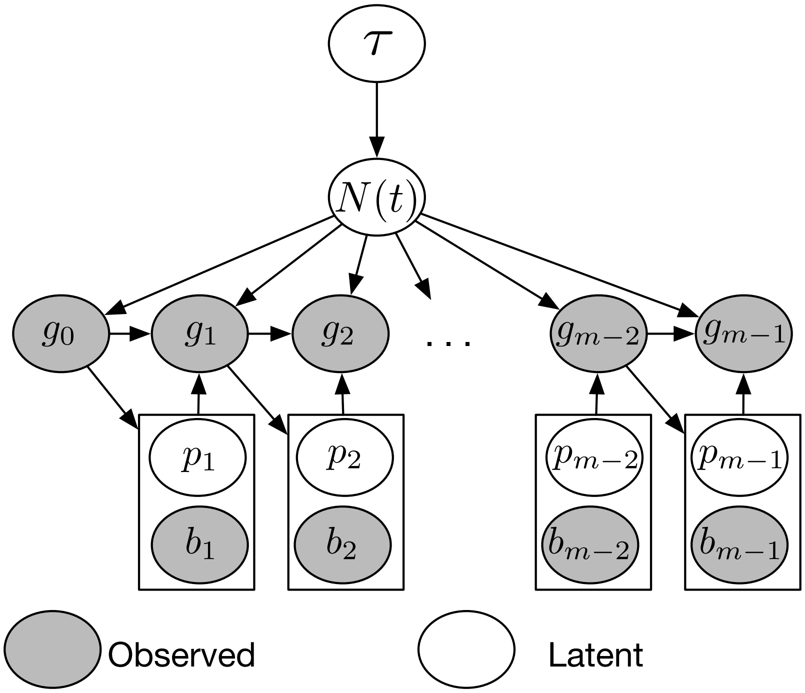


Figure 3: **Structure of our Bayesian model for inferring population size trajectories from a realization of the SMC′ process at recombination breakpoints**. Hyperparameter τ controls the smoothness of the log-Gaussian process prior on $N(t)$. Local genealogies depend on $N(t)$ and form a Markov chain of degree one. Given the current local genealogy g_{i-1}, we sample the location of the new recombination breakpoint b_i and a pruning location p_i on genealogy g_{i-1}. The new genealogy g_i depends on $N(t)$, p_i and g_{i-1}.

3.2 Maximum-likelihood estimation of $N(t)$ with measures of uncertainty

We assume that the population size trajectory $N(t)$ is defined as in Equation 14. The standard coalescent density (Equation 4) and the transition densities defined in Equations 11 and 8 are tractable, so calculation of the likelihood (Equation 13) is tractable. However maximization of the likelihood function cannot be performed analytically because pruning locations are missing. We rely on the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) to find the maximum likelihood estimator of $\mathbf{N} = (N_1, \ldots, N_d)$. The complete data $\mathbf{Y}_c$ for inferring $N(t)$ are then the set of local genealogies $g_0, \ldots, g_{m-1}$ and the set of pruning locations $p_1 \ldots, p_{m-1}$. For the invisible transitions, we also need to know the new coalescent times $\{t^i_{new}\}_{i \in \mathcal{I}}$, where $\mathcal{I} \subset \{1, 2, \ldots, m-1\}$ denotes the set of indices of invisible transitions (transition i is an invisible transition if $g_i = g_{i-1}$).

The complete data log-likelihood is then

$$\mathcal{L}_c(\mathbf{Y}_c; \mathbf{N}) := \log \Pr[g_0 \mid N(t)] + \sum_{i=1}^{m-1} \log \Pr[p_i = (u_i, w_i), t^i_{new}, c_i \mid g_{i-1}, N(t)]. \tag{20}$$

The EM algorithm starts by initializing the population size trajectory to a piece-wise constant function with change points $x_1, \ldots, x_d$ with arbitrarily chosen vector $\mathbf{N}^0$. At the kth iteration of the algorithm we set

$$\mathbf{N}^k = \arg\max_{\mathbf{N}} \mathrm{E}_{\mathbf{N}^{k-1}}[\mathcal{L}_c(\mathbf{Y}_c; \mathbf{N}) \mid \mathbf{Y}]. \tag{21}$$

The conditional expectation in Equation 21 is conditional on the observed data $\mathbf{Y}$ defined in Equation 12. Let $\mathbf{x}^i = \{x^i_1, x^i_2, \ldots, x^i_{d+n-1}\}$ be the ordered set of time points corresponding to the change points $x_1, \ldots, x_d$ and the coalescent time points $\mathbf{t}^i$ of local genealogy i. If the transition from g_i to g_{i+1} is visible, we replace the jth time point x^i_j by t^{i+1}_{new}, where j corresponds to the index such that $x^i_{j-1} < t^{i+1}_{new} \leq x^i_j$. For ease of notation, we will denote the number of time intervals $|\mathbf{x}^i|$ by $D = d + n - 2$. Let

$$a^0_j = \begin{cases} 1, & \text{if } x^0_{j+1} = t^0_k, \text{ for } k = 2, \ldots, n, \\ 0, & \text{othewise,} \end{cases}$$

be an indicator function that takes the value of 1 when the jth interval contains a coalescent time of the first genealogy g_0. Then, the log density of the first genealogy is:

$$\log \Pr[g_0 \mid N(t)] = -\sum_{j=1}^{D} \left\{ a^0_j logN(x^0_{j+1}) + \frac{A^0(x^0_{j+1})[A^0(x^0_{j+1}) - 1](x^0_{j+1} - x^0_j) \exp[-\log N(x^0_{j+1})]}{2} \right\}. \tag{22}$$

Let

$$z^i_j = \begin{cases} 1, & \text{if } x^i_j < t^{i+1}_{new} \leq x^i_{j+1}, \\ 0, & \text{otherwise,} \end{cases}$$

be an indicator function that takes the value of 1 when the new coalescent time of genealogy i

happens in the corresponding time interval (x^i_j, x^i_{j+1}), and let the adjusted interval length be

$$\Delta^i_j = \begin{cases} x^i_{j+1} - x^i_j, & \text{if } u_{i+1} < x^i_j, \text{ and } x^i_{j+1} < t^{i+1}_{new} \text{ (after pruning and before coalescence)}, \\ x^i_{j+1} - u_{i+1}, & \text{if } x^i_j < u_{i+1} < x^i_{j+1} \leq t^{i+1}_{new} \text{ (before coalescence with pruning adjustment)}, \\ t^{i+1}_{new} - u_{i+1}, & \text{if } x^i_j < u_{i+1} < t^{i+1}_{new} < x_{j+1} \text{ (adjustment for prunning and coalescence)}, \\ t^{i+1}_{new} - x^i_j, & \text{if } u_{i+1} < x^i_j < t^{i+1}_{new} < x^i_{j+1} \text{ (after pruning with coalescence adjustment)}, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the augmented transition density can be expressed as:

$$\begin{aligned} \log \Pr[p_i = (u_i, w_i), t^i_{new}, c_i \mid g_{i-1}, N(t)] &= \log \Pr[p_i = (u_i, w_i), t^i_{new}, c_i, \mathbf{z}^i, \boldsymbol{\Delta}^i \mid g_{i-1}, N(t)] \\ &= -\log l_{i-1} - \sum_{j=1}^{D} \left\{ z^{i-1}_j \log N(x^{i-1}_{j+1}) \right\} \\ &\quad - \sum_{j=1}^{D} \left\{ A^{i-1}(x^{i-1}_{j+1}) \Delta^{i-1}_j \exp[-\log N(x^{i-1}_{j+1})] \right\}. \qquad (23) \end{aligned}$$

where $\mathbf{z}^i$ and $\boldsymbol{\Delta}^i$ are the vectors with z^i_j and Δ^i_j elements. For the EM algorithm we need to compute the conditional expected vectors $\mathrm{E}[\mathbf{z}^i_j \mid \mathbf{Y}]$ and $\mathrm{E}[\boldsymbol{\Delta}^i_j \mid \mathbf{Y}]$. The details of these calculation are in the Appendix.

We use the Fisher information matrix to compute approximate standard errors of $\log \widehat{\mathbf{N}}$ and use these standard errors together with asymptotic normality of maximum likelihood estimators to produce confidence intervals for log population size piece-wise trajectories. We compute the observed Fisher information matrix following Louis (1982):

$$\hat{\mathbf{I}}_{\mathbf{Y}}[\widehat{\mathbf{N}}] = \mathrm{E}_{\widehat{\mathbf{N}}}[-\mathbf{H}\mathcal{L}_c(\mathbf{Y}_c; \widehat{\mathbf{N}}) \mid \mathbf{Y}] - \mathrm{E}_{\widehat{\mathbf{N}}}[\nabla \mathcal{L}_c(\mathbf{Y}_c; \widehat{\mathbf{N}}) \nabla \mathcal{L}_c(\mathbf{Y}_c; \widehat{\mathbf{N}})' \mid \mathbf{Y}],$$

where $\nabla \mathcal{L}_c(\mathbf{Y}_c; \widehat{\mathbf{N}})$ is the gradient and $\mathbf{H}\mathcal{L}_c(\mathbf{Y}_c; \widehat{\mathbf{N}})$ is the Hessian of the complete-data log-likelihood with respect to $\log \mathbf{N}$. This requires the calculation of conditional cross-product means and conditional second moments described in the Appendix.

4 Results

We simulated 1000 local genealogies of 2, 20 and 100 individuals from each of the three different demographic models described in Table 2 using `MaCS` (Chen et al. 2009); see Supporting Information for details of these simulations. We assumed that all individuals were sampled at time $t = 0$ under a demographic model in Table 2.

We compared the point estimates with the truth for each demographic model using the sum of relative errors (SRE):

$$\mathrm{SRE} = \sum_{i=1}^{K} \frac{|\widehat{N}(x_i) - N(x_i)|}{N(x_i)}, \qquad (24)$$

Table 2: Simulated demographic scenarios. The argument t denotes time measured in units of N_0 generations.

Demographic model	$N(t)$
Constant Population size:	$N(t) = 1$
Exponential growth followed by constant size:	$N(t) = \begin{cases} 1, & \text{for } t \in (0, 0.1), \\ \exp[-10(t-0.1)], & \text{for } t \in (0.1, \infty). \end{cases}$
Population bottleneck:	$N(t) = \begin{cases} 1, & \text{for } t \in (0, 0.3), \\ 0.1, & \text{for } t \in (0.3, 0.5), \\ 1, & \text{for } t \in (0.5, \infty). \end{cases}$

where $\widehat{N}(x_i)$ is the estimated population size trajectory at time x_i. We compute SRE at equally space time points $x_1, \ldots, x_K$. Second, we compute the mean relative width (MRW) as follows:

$$\text{MRW} = \sum_{i=1}^{K} \frac{|\widehat{N}_{up}(x_i) - \widehat{N}_{low}(x_i)|}{KN(x_i)}, \tag{25}$$

where $\widehat{N}_{up}(x_i)$ corresponds to the 97.5% upper limit and $\widehat{N}_{low}(x_i)$ corresponds to the 2.5% lower limit of $\widehat{N}(x_i)$. For EM estimates, $[\widehat{N}_{low}(x_i), \widehat{N}_{up}(x_i)]$ corresponds to the 95% confidence interval estimated using the observed Fisher information; for Bayesian GP estimates, $[\widehat{N}_{low}(x_i), \widehat{N}_{up}(x_i)]$ corresponds to the 95% Bayesian credible interval (BCI) of $\widehat{N}(x_i)$. To measure how well these intervals cover the truth, we compute the envelope measure (ENV) in the following way:

$$\text{ENV} = \frac{\sum_{i=1}^{K} I(\widehat{N}_{up}(x_i) \leq N(x_i) \leq \widehat{N}_{low}(x_i))}{K} \tag{26}$$

We compute SRE, MRW and ENV for $K = 150$ at equally spaced time points.

For our Bayesian GP estimates, we estimate $N(x_i)$ at $d = 100$ time points, unless stated otherwise. The parameters of the Gamma prior on the GP precision parameter τ were set to $\alpha = \beta = 0.001$, reflecting our lack of prior information about the smoothness of the population size trajectory.

For our EM estimates, we used different discretizations based on Equation 15 and varying the number of change points d and κ over the fixed interval $(0, x_d)$ with x_d set to be the maximum observed coalescent time. For the cases where we only consider one genealogy ($m = 1$), the EM approach becomes standard maximum likelihood estimation. We summarize our posterior inference and compare our Bayesian GP method to the EM method. The population size trajectory is log-transformed for ease of visualization and for direct comparison with other methods (Minin et al. 2008; Palacios and Minin 2013).

Table 3: Summary statistics for simulation results depicted in Figure 4. SRE is the sum of relative errors (Equation 24), MRW is the mean relative width of the 95% BCI (Equation 25), and ENV is the envelope measure (Equation 26). Values in bold indicate best performance.

Simulation of a single genealogy with $n = 100$

	SRE	MRW	ENV
MLE $d = 5$, $\kappa = 1$	41.80	14.76	**100.0%**
MLE $d = 5$, $\kappa = 10$	41.05	2.98	**100.0%**
MLE $d = 5$, $\kappa = 100$	57.12	1.72	**100.0%**
MLE $d = 10$, $\kappa = 10$	47.93	16.08	**100.0%**
MLE $d = 10$, $\kappa = 100$	61.77	3.91	**100.0%**
MLE $d = 10$, $\kappa = 500$	31.52	3.60	**100.0%**
Bayesian GP $d = 50$	6.98	1.88	**100.0%**
Bayesian GP $d = 100$	5.52	2.15	**100.0%**
Bayesian GP $d = 200$	**4.96**	**1.70**	**100.0%**

4.1 Sensitivity of EM estimates of $N(t)$ to discretization

In Figure 4, we show our Bayesian GP and EM estimates of a constant population size trajectory from a single genealogy of 100 individuals with different discretizations. We find that our Bayesian GP point estimates depicted in Figure 4A recover the truth (dashed line) almost perfectly with less uncertainty than the EM (Figure 4B-C). Comparing our Bayesian GP estimates with different discretizations: 50, 100 and 200 equally spaced time points (Figure 4A), we find that increasing the number of time points improves inference (Table 4) but that the differences between estimates among the three discretizations are marginal (Figure 4A). In contrast, we show that different grid definitions alter the EM estimates (Figure 4B). It is not clear how to define a good strategy for the definition of the grid for the EM method, even for the simple model of constant population size. For example, increasing κ from 100 to 500 with 5 change points (Figure 4B), does not improve estimation. Increasing the number of change points does not necessarily improve the estimates either; for example, increasing the the number of change points from 5 to 10 for $\kappa = 10$ (Figures 4B-C). EM grid sensitivity is persistent even when the number of genealogies increases; Figure S2 in Supplementary Information shows that the best definition of change points when our data consist of 1000 local genealogies of 100 individuals has 10 change points evenly distributed.

4.2 Comparing Methods of Estimating $N(t)$

Figure 5 shows the estimated population size trajectories when the number of samples is 2 for the three different demographic scenarios and varying the number of local genealogies (100, 500 and 1000 local genealogies). For constant and exponential growth, our EM method assumes a piece-wise constant trajectory of 10 change points ($d = 10$) and $\kappa = 1$ using Equation 15 (similar to Li and Durbin (2011) and Rasmussen et al. (2014)). For the bottleneck scenario, some of the intervals did not have coalescent events; hence, for this case we assumed a piece-wise constant trajectory of 5

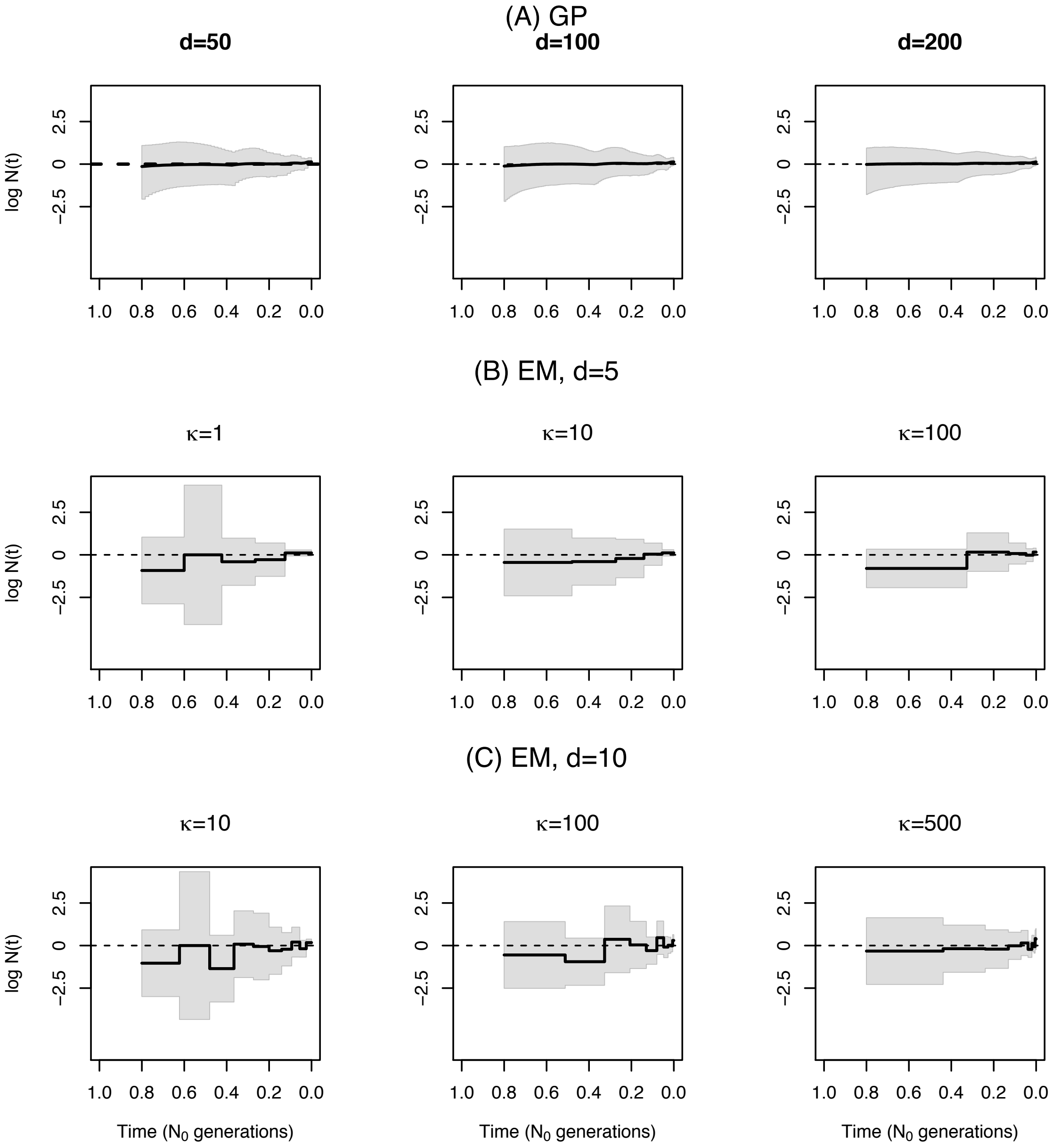


Figure 4: **Sensitivity to parameter discretization.** Comparison of population size trajectories estimated from one simulated genealogy ($m = 1$) of 100 individuals with a constant population size. We show true trajectories as dashed lines. (A) Bayesian GP estimates at $d = 50, 100$ and 200 equally spaced time points. (B) EM estimates of a piece-wise constant trajectory with $d = 5$ change points and $\kappa = 1, 10$ and 100 (Equation 15). (C) EM estimates of a piece-wise constant trajectory with $d = 10$ change points and $\kappa = 10, 100$ and 500 (Equation 15). Point estimates are shown as solid black lines. 95% credible intervals and 95% confidence intervals are shown by gray shaded areas.

change points ($d = 5$) and $\kappa = 1$ for constructing our EM estimates. We show the boxplots of the time to the most recent common ancestor (TMRCA) at the bottom of each plot in Figure 5. The distribution of the TMRCA serves as an indicator of the uncertainty expected of our estimates. Both approaches, EM and Bayesian GP show narrower confidence and credible intervals at the center of the distribution of the TMRCA, particularly during the bottleneck in Figure 5C.

For the constant population demographic model in Figure 5A, our Bayesian GP outperforms our EM estimates considerably. This is not surprising since *a priori* $\log N(t)$ has mean 0 in our Bayesian approach (Equation 17). Moreover, EM confidence intervals only cover the truth constant population size around 30% of the time, while the GP method covers 100% of the truth (Table 4A). Despite placing a mean-0 prior on $logN(t)$, the Bayesian GP method accurately recovers sudden changes as shown in the bottleneck example. Our Bayesian GP prior on $\log N(t)$ is Brownian motion which is not differentiable at any point; yet, our Bayesian GP recovers smooth curves (Figure 5B).

Table 4A shows the performance statistics for the estimates of $N(t)$ in Figure 5. In general, our Bayesian GP has wider credible intervals than the EM confidence intervals but these credible intervals cover the true trajectory better than the EM confidence intervals in all the cases (MRW and ENV in Table 4). Our Bayesian GP estimates also generally have smaller sums of relative errors (SRE in Table 4). Under the bottleneck scenario, our Bayesian GP produces greater sums of relative errors than does the EM, but our Bayesian GP estimates recover the truth more accurately than the EM during the bottleneck.

Figures 6 and 7 show our estimates when $n = 20$ and $n = 100$. The performance statistics of the estimates displayed are shown in Table 4(B) and (C). In general, our GP-based estimates have smaller SRE and larger ENV than the EM-based estimates and hence, the MRW is usually wider in the GP-based estimates, accurately reflecting the uncertainty of the estimates. As expected, increasing the number of loci (m) generally decreases the width of the confidence and credible intervals of our estimates (MRW). Although this is generally true for EM estimates as well, EM estimates have very low coverage of the truth (MRE in Table 4) when the number of loci increases.

4.3 Sampling more individuals versus sequencing more loci

Figures 5-7 show our estimates for $n = 2, 20$ and 100 sampled individuals across varying numbers of loci. Since performance of EM estimates depends strongly on the definition of the grid, we base what follows on the Bayesian GP estimates. We find that increasing the number of loci, decreases uncertainty of our estimates and allows us to infer $N(t)$ further back in time. Increasing the number of samples does not necessarily increase the performance of our GP estimates. For example, under the bottleneck scenario, we are able to detect the bottleneck phase fairly accurately even for two samples with $m = 1000$ local genealogies. Increasing the number of samples to $n = 20$ and $n = 100$ does not improve estimation of the features of the bottleneck. This is because most TMRCAs observed under the bottleneck scenario occur during the bottleneck (Figures 5,6 and 7), regardless of the number of individuals sampled. In contrast, in our exponential growth example, increasing the number of samples from $n = 2$ to $n = 100$ improves accuracy (point estimates are closer to the truth, see SRE in Tables 4A-C) and credible intervals cover the truth completely (ENV of 100%).

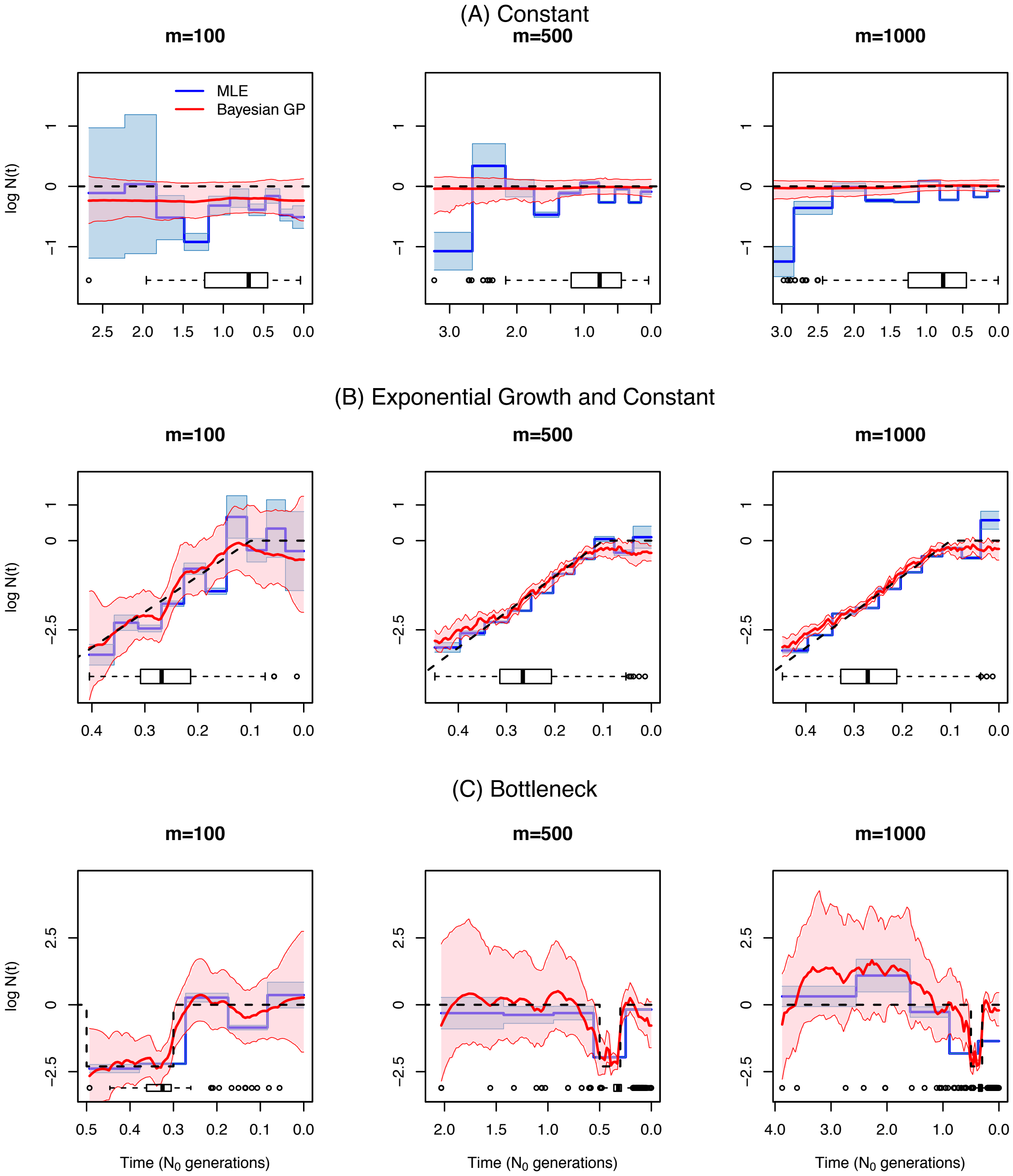


Figure 5: **Inference of population size trajectories $N(t)$ for a pair of individuals** $(n = 2)$. (A) Simulated data under constant population size, (B) exponential and constant trajectory, and (C) a bottleneck. We show estimates from $m = 100$, $m = 500$, and $m = 1000$ local genealogies. We show the true trajectories as dashed lines, blue lines and light blue shaded areas represent EM point estimates and 95% confidence areas, and red lines and pink shaded areas represent Bayesian GP posterior medians and 95% BCIs. Boxplots of the TMRCA are shown at the bottom of each plot.

Table 4: Summary of simulation results depicted in Figures 5. SRE is the sum of relative errors calculated as in (24), MRW is the mean relative width of the 95% BCI as defined in (25), and ENV is the envelope measure calculated as in (26). Values in bold indicate best performance for each demographic model and sample size.

A. Simulations with $n = 2$

	SRE			MRW			ENV		
	m=100	m=500	m=1000	m=100	m=500	m=1000	m=100	m=500	m=1000
Const. EM	39.80	41.78	38.60	0.98	**0.26**	**0.08**	31.3%	28.0%	19.3%
Const. GP	**30.60**	**4.25**	**3.04**	**0.49**	0.33	0.22	**100.0%**	**100.0%**	**100.0%**
Exp. EM	64.68	**25.70**	33.70	**0.91**	**0.16**	**0.12**	42.0%	26.0%	6.6%
Exp. GP	**28.38**	32.70	**26.76**	2.04	0.45	0.33	**100.0%**	**56.0%**	**50.6%**
Bottle. EM	48.48	46.51	**127.70**	**0.43**	**0.45**	**1.37**	40.6%	30.0%	34.0%
Bottle. GP	**33.76**	**45.14**	223.58	3.44	6.84	17.13	**98.0%**	**94.6%**	**94.6%**

B. Simulation with $n = 20$

	SRE			MRW			ENV		
	m=1	m=100	m=1000	m=1	m=100	m=1000	m=1	m=100	m=1000
Const. EM	60.87	121.30	25.60	2.28	2.16	0.23	**100.0%**	37.7%	39.3%
Const. GP	**31.74**	**3.94**	**13.22**	**1.06**	**0.70**	0.36	**100.0%**	**100.0%**	**100.0%**
Exp. EM	40.97	40.66	**40.22**	**3.11**	**0.37**	**0.19**	**100.0%**	38.6%	19.3%
Exp. GP	**25.35**	**27.03**	65.61	3.53	1.56	0.42	**100.0%**	**100.0%**	**39.3%**
Bottle. EM	147.93	78.40	78.20	6.98	**0.81**	68.4	66.0%	78.6%	49.33%
Bottle. GP	**68.93**	**78.2**	**50.92**	**2.74**	2.47	**1.47**	**92.0%**	**79.3%**	**78.6%**

C. Simulation with $n = 100$

	SRE			MRW			ENV		
	m=1	m=100	m=1000	m=1	m=100	m=1000	m=1	m=100	m=1000
Const. EM	41.05	220.85	43.41	2.98	4.93	0.99	**100.0%**	35.3%	48.0%
Const. GP	**5.52**	**34.78**	**12.17**	**2.15**	**1.49**	**0.47**	**100.0%**	**100.0%**	**89.3%**
Exp. EM	**76.86**	40.22	27.63	**3.23**	**0.81**	**0.13**	87.3%	42.0%	14.0%
Exp. GP	114.53	**25.82**	**26.42**	3.57	1.55	0.83	**100.0%**	**100.0%**	**100.0%**
Bottle. EM	194.77	59.54	127.68	**3.95**	**1.08**	**0.85**	84.0%	51.3%	45.3%
Bottle.	**90.27**	**44.14**	**42.68**	6.98	2.62	1.74	**100.0%**	**94.7%**	**96.0%**

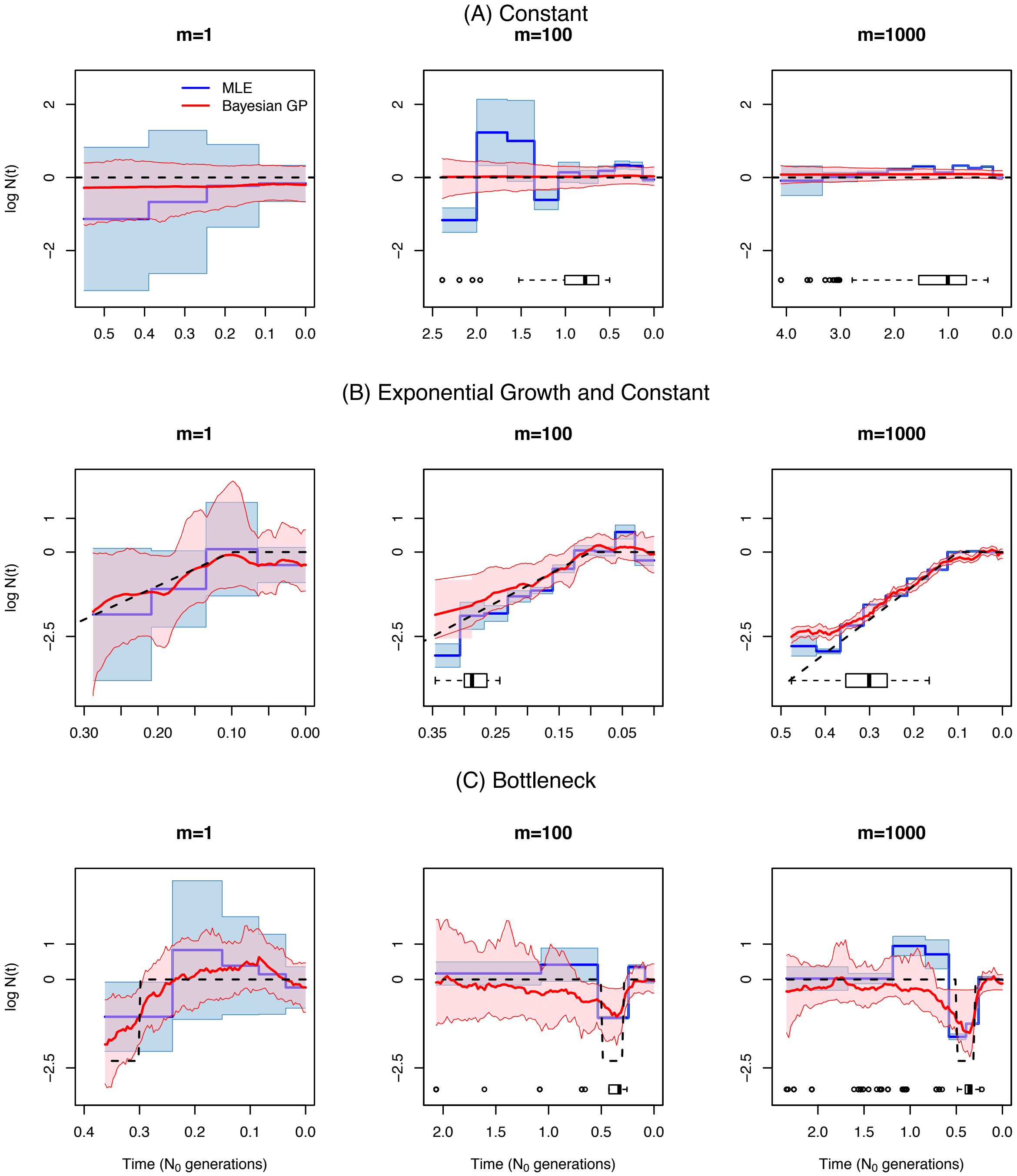


Figure 6: **Inference of population size trajectories $N(t)$ for $n = 20$.** (A) Simulated data under constant population size, (B) exponential and constant trajectory, and (C) a bottleneck. We show estimates from $m = 1$ genealogy, $m = 100$ local genealogies and $m = 1000$ local genealogies. We show the true trajectories as dashed lines, blue lines and light blue shaded areas represent EM point estimates and 95% confidence areas, and red lines and pink shaded areas represent Bayesian GP posterior medians and 95% BCIs. Boxplots of the TMRCA are shown at the bottom of each plot.

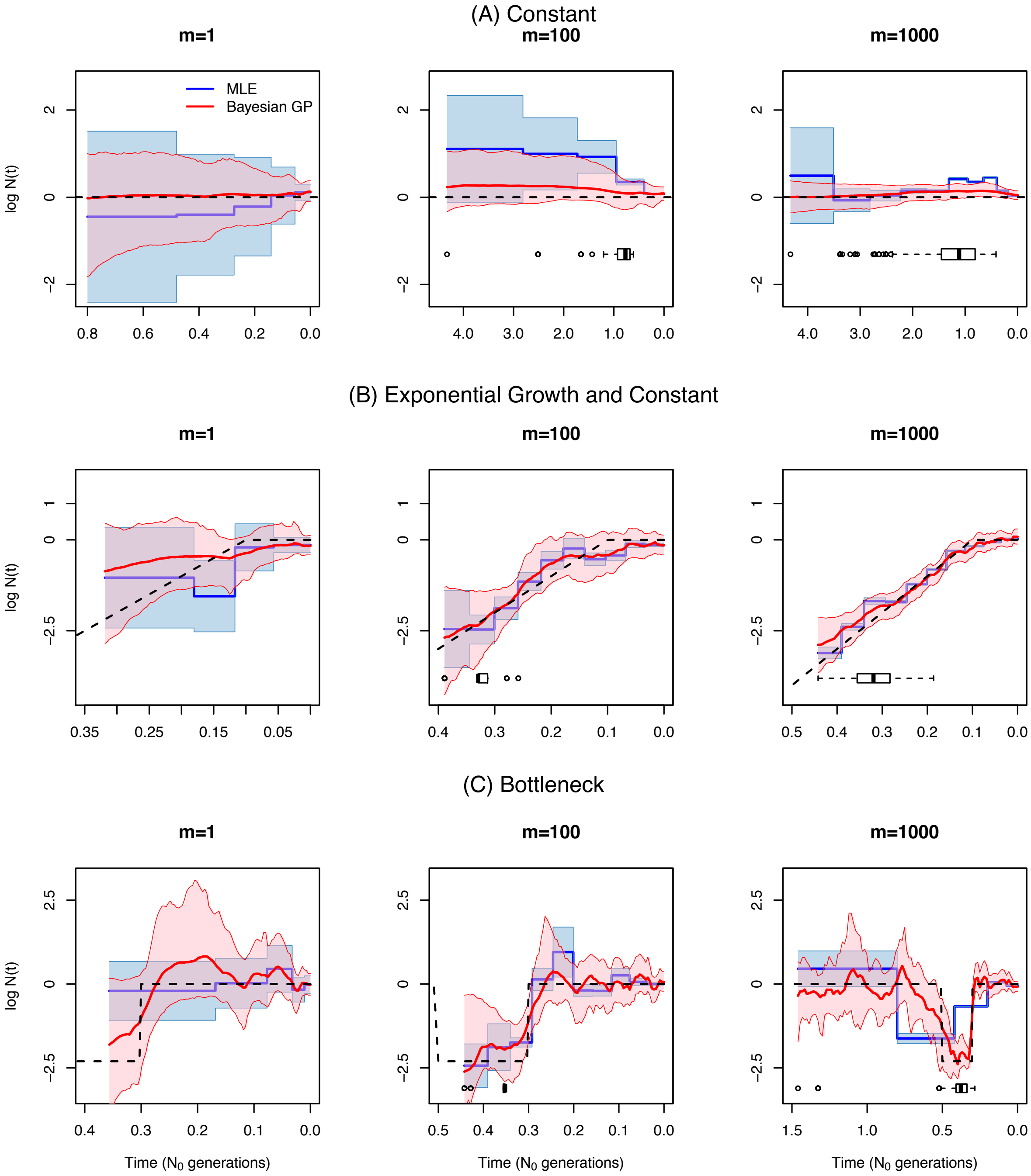


Figure 7: **Inference of population size trajectories $N(t)$ for $n = 100$.** (A) Simulated data under constant population size, (B) exponential and constant trajectory, and (C) bottleneck. We show estimates from $m = 1$ genealogy, $m = 100$ local genealogies and $m = 1000$ local genealogies. We show the true trajectories as dashed lines, blue lines and light blue shaded areas represent EM point estimates and 95% confidence areas, and red lines and pink shaded areas represent Bayesian GP posterior medians and 95% BCIs. Boxplots of the TMRCA are shown at the bottom of each plot.

4.4 Sequential Tajima's genealogies are sufficient statistics under the SMC′

Under the SMC′ process, marginally at each locus along the chromosome, a local genealogy is a realization of Kingman's n-coalescent (Kingman 1982), a continuous-time Markov chain taking its values in the set $\mathcal{K}_n$ of partitions of the label set $\{1, 2, \ldots, n\}$. A local genealogy g of n individuals includes labeled topology K_n and coalescent times $\mathbf{t} = (t_n, \ldots, t_2)$. The state space of a local genealogy is then $\mathcal{G} = \mathcal{K}_n \otimes \mathbb{R}^{+n-1}$, and the cardinality of the set $\mathcal{K}_n$ is $n!(n-1)!/2^{n-1}$. However, only the set of ordered coalescent times carry information about $N(t)$. For a single locus, the set coalescent times are sufficient statistics for inferring $N(t)$ (proof is in the Appendix). A natural question that follows is whether the coalescent times corresponding to the set of local genealogies are sufficient statistics for inferring $N(t)$ under the SMC′ model. We find that the sufficient statistics for inferring $N(t)$ under the SMC′ model, are the coalescent times, when taken together with local ranked tree shapes. For a single locus, the set of coalescent times together with the ranked tree shape correspond to a realization of Tajima's n-coalescent. Tajima's n-coalescent (Tajima 1983) is a continuous-time Markov chain taking its values in the set $\mathcal{H}_n$ of ranked tree shapes also called histories, evolutionary relationships or vintaged and sized coalescent (Sainudiin et al. 2014). The state space of Tajima's local genealogy is then $\mathcal{G}^T = \mathcal{H}_n \otimes \mathbb{R}^{+n-1}$, and the cardinality of the set $\mathcal{H}_n$ corresponds to the sequence of Euler zigzag numbers whose first ten elements are $1, 1, 1, 2, 5, 16, 61, 272, 1385, 7936$ (Disanto and Wiehe 2013). The probability of getting a particular type of ranked tree shape H_n of n samples (Tajima 1983) is given by

$$P(H_n) = \frac{2^{n-c-1}}{(n-1)!}, \tag{27}$$

where c is the number of *cherries*, defined as branching events that lead to exactly two leaves.

In the Methods section, we defined transition densities in terms of coalescent times and $F_{i,j}$ quantities. The set of all $F_{i,j}$ quantities from a local genealogy form a triangular matrix: F-matrix. In the Appendix, we show that (i) F-matrices are in bijection with ranked tree shapes and (ii) the set of local Tajima's genealogies are sufficient statistics for inferring $N(t)$ under the SMC′ model. These observations are crucial for inferring $N(t)$ from sequence data directly. Coalescent-based inference from sequence data rely on marginalization over the hidden state space of genealogies. In the Appendix, we show that the state space needed is the space of local Tajima's genealogies, as opposed to the space of local Kingman's genealogies. For $n = 10$ sequences, there are $2,571,912,000$ possible labeled topologies while only $7,936$ possible ranked tree shapes.

4.5 Application to human data

We applied our method to a 2-Mb region on chromosome 1 (187,500,000-189,500,000) with no genes from five Yorubans from Ibadan, Nigeria (YRI) and five Utah residents of central European descent (CEU) from the 1000 Genomes pilot project (1000 Genomes Project Consortium 2012) and previously analyzed for the same purpose (Sheehan et al. 2013). We used *ARGweaver* (Rasmussen et al. 2014) to obtain a sample path of local genealogies for the two populations (YRI and CEU). The parameters used are 200 change points, a mutation rate of $\mu = 1.26 \times 10^{-8}$ and a recombination rate of $\rho = 1.6 \times 10^{-8}$ (Rasmussen et al. 2014, details regarding parameters used can be found in

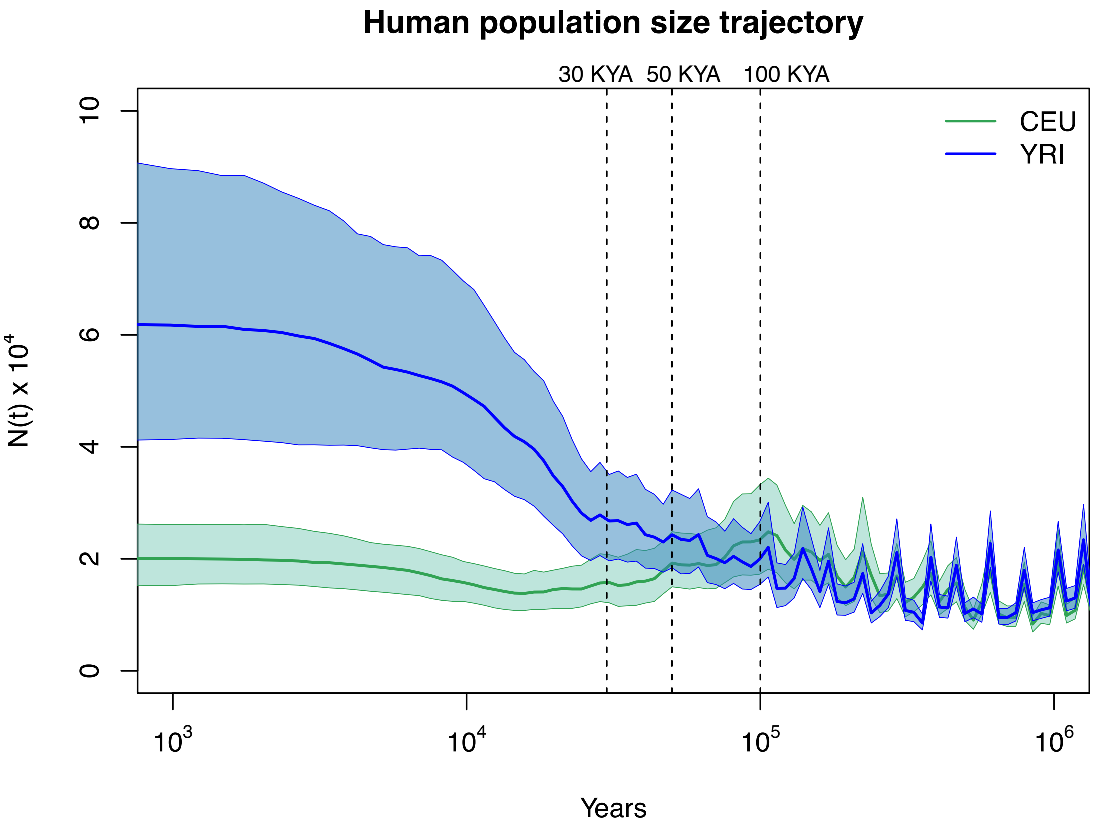


Figure 8: **Inference of human population size trajectories $N(t)$ for $n = 10$.** Green solid line and green shaded areas represent the posterior median and 95% BCI for European population (CEU) and blue solid line and blue shaded areas represent the posterior median and 95% BCI for Yoruban population (YRI). Time is measured in years in the past assuming a generation length of 25 years and a reference diploid population of 10,000 individuals. The x-axis is log transformed.

Supplementary Information). We note that *ARGweaver* assumes the SMC process and our method assumes the SMC$'$ process. Moreover, our inference is based on a single sample of the SMC process with known pruning times. Our *ARGweaver* set of local genealogies are discretized at 200 time points and our GP-based inference is influenced by this discretization. In Figure 8 we show our estimates of past Yoruban (in blue) and European population sizes (in green). The two population size trajectories experience a series of bottlenecks and overlap until about 100 YKA, assuming a diploid reference population size of N_0=10,000 and a generation time of 25 years. In Figure 8 we recover an out-of-Africa bootleneck that starts about 100 KYA and ends about 30 KYA in the European population. These results are consistent with previously published results (Li and Durbin 2011; Gronau et al. 2011; Rasmussen et al. 2011; Sheehan et al. 2013; Schiffels and Durbin 2014). In Supplementary Information Figure S4 we show the estimates of $logN(t)$ instead of $logN(t)$ and time measured in units of N_0 generations (same scaling as with simulations in Figures 5-7).

5 Discussion

In this paper, we propose a Gaussian-process based Bayesian nonparametric method for estimating effective population size trajectories $N(t)$ from a sequence of local genealogies, accounting for recombination. Under a variety of simulated demographic scenarios and sampling designs, our method recovers the truth with better precision and accuracy than a maximum likelihood approach (Figures 5-7). We apply our method to genealogies estimated from human genomic data *ARGweaver* (Rasmussen et al. 2014) and conduct inference of the human population size trajectory for European and African populations; this application to real data recover the known features of the out of Africa bottleneck (Figure 8).

Several recent approaches have emerged for inference of population size trajectories from multiple whole-genome sequences using the sequentially Markov coalescent (SMC) (Li and Durbin 2011; Sheehan et al. 2013; Schiffels and Durbin 2014). However, current SMC-based methods rely on maximum likelihood inference (EM) of both a discretized parameter space and a discretized state space in order to gain computational tractability, and incur the costs of reduced accuracy and biased estimates. Although in principle the EM approach and the Bayesian nonparametric approach approximate $N(t)$ similarly — by either a piece-wise constant or a piece-wise linear function — the Bayesian nonparametric approach is not affected by increasing the number of parameters (or change points) in the estimation of $N(t)$. For comparison with existing methods, we implemented an EM approach to infer population size trajectories from a sequence of local genealogies and we note that increasing the number of loci may actually increase the bias of the EM estimates (Figures 5-7). For example, in simulation, our EM approach incorrectly detects the initial period of the simulated bottleneck (around $0.8N_0$ instead of $0.5N_0$ generations ago) with narrow confidence intervals (Figure 7C).

There are many advantages to using Bayesian GP over EM for inference of population size trajectories. Similar to Palacios and Minin's (2013) approach to inference from a single genealogy, we a *priori* assume that $N(t)$ follows a log Brownian Motion process. This allows us to model $N(t)$ as a continuous positive function. The main advantage of using a Brownian Motion process is that its inverse covariance function is a sparse matrix that allows for fast computations. Since the likelihood function involves integration over $N(t)$, this integral is approximated by the Riemann sum over a regular grid of points. The finer the grid is, the better the approximation. We find that our method performs well for inferring $N(t)$ at 100 change points in all our examples and, more importantly, results are not sensitive to the number of change points used in the analysis (Figure 4). Our Bayesian approach relies on MCMC for inference from the posterior distribution of model parameters. Because population sizes at different grid points are correlated, we adapt the recently developed MCMC technique Split Hamiltonian Monte Carlo (splitHMC) for jointly sampling all model parameters (Shahbaba et al. 2014; Lan et al. 2015). splitHMC is a Metropolis sampling algorithm that efficiently proposes states that are distant from current states with high acceptance rates. It has been shown to be more efficient in inferring $N(t)$ from a single genealogy than elliptical slice sampling or regular Hamiltonian Monte Carlo sampling(Lan et al. 2015). However, splitHMC relies on calculating the score function at every single iteration. Because pruning time in each local genealogy is unknown, we calculate the score function via Fisher's formula.

In simulations, we find that our algorithm scales well with hundreds of individuals; our computational bottleneck is in the number of local genealogies. We envision that extending the current methodology to inference from sequence data directly will require a strategy for sampling shorter genomic segments. This would be a probabilistic alternative to arbitrarily choosing segment lengths (Sheehan et al. 2013; Rasmussen et al. 2014).

Under the SMC model, every recombination event along the genome translates to a new coalescent event for the sample under study, so increasing the number of loci results in more realizations of the coalescent process. The longer the segments are and the larger the number of samples taken, the greater the chance of observing variation due to recombination. This fact makes it hard to define a sampling strategy: longer genomes or larger sample sizes? We show that increasing the number of local genealogies improves precision of our Bayesian GP estimates (Figures 5-7). However, resolution into the past from contemporaneous sequences highly depends on the actual population size trajectory $N(t)$.

We use *ARGweaver* (Rasmussen et al. 2014) to generate two samples of contiguous local genealogies corresponding to a 2-Mb region of chromosome 1 for five Europeans (CEU) and five Africans (YRI) from the 1000 Genomes Project; this genomic region is free of genes and was also analyzed in Sheehan et al. (2013). Taking these two samples of local genealogies as our data (4186 local genealogies for CEU and 6247 local genealogies for YRI), we were able to use our Bayesian GP method to infer Yoruban and and European effective population size trajectories (Figure 8). We find an out-of-Africa bottleneck that began $\sim$ 100 KYA and ended $\sim$ 30 KYA in the European population consistent with Li and Durbin (2011); Rasmussen et al. (2011); Gronau et al. (2011); Sheehan et al. (2013) and Schiffels and Durbin (2014). We note that our estimates are based on a single sample of local genealogies and thus ignore genealogical uncertainty. Moreover, we generated our data from the posterior distribution of local genealogies using *ARGweaver* at 200 time intervals so our GP-based approach cannot fully detect sudden changes that may occur between the discretized times. In addition, *ARGweaver* assumes an SMC prior model on local genealogies and our GP-based method assumes the SMC$'$ process; the lack of invisible recombination events in *ARGweaver*'s genealogies will bias inference.

The natural next extension for our method presented in this study is to infer $N(t)$ from sequence data directly and not from the set of local genealogies. Our MCMC approach allows us extend the current methodology in a Bayesian hierarchical framework where the SMC$'$ process would be used as a prior distribution over local genealogies. The work we present here suggests a combination of *ARGweaver* accommodating SMC$'$ and GP priors would result in an efficient method for inferring population size trajectories from sequence data directly. In addition, our model can be easily modified to model a variable recombination rate along chromosomal segments and to jointly infer variable recombination rates and $N(t)$.

Finally, we show that, under the SMC$'$ model, local ranked tree shapes and coalescent times correspond to a set of local Tajima's genealogies; these Tajima's genealogies are the sufficient statistics for inferring $N(t)$. Under the SMC$'$ model, the state space needed for inferring population size trajectories from sequence data is that of a sequence of local Tajima's genealogies. This lumping, or reduction of the original SMC′ process, will allow more efficient inference from sequence data directly.

Current methods for inferring population size trajectories make tradeoffs to analyze whole genomes that limit both biological understanding of sudden population size changes and the ability to test hypotheses regarding population size changes. This work represents a critical set of theoretical results that lay the groundwork for efficient estimation of detailed histories from sequence data with measures of uncertainty.

Acknowledgements

We thank Amandine Veber for her valuable suggestions and comments. We thank Shiwei Lan for his suggestions for speeding up the MCMC sampling scheme and Sara Sheehan and Melissa J. Hubisz for helpful discussions. J.A.P. acknowledges scholarship from CONACyT Mexico to pursue her research work. This research is supported in part by NSF CAREER Award DBI-1452622 (to S.R.). S.R. is a Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts, and an Alfred P. Sloan Research Fellow.

Literature Cited

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56 – 65.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statitistical Society. Series B*, 39(1):1–38.

Disanto, F. and Wiehe, T. (2013). Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. *Mathematical Biosciences*, 242(2):195 – 200.

Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. In Donnelly, P. and Tavaré, S., editors, *Progress in population genetics and human evolution*, volume 87 of *IMA Volumes in Mathematics and Its Applications*, pages 257–270. Springer Verlag, New York.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, 43(10):1031–1034.

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37:203–217.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–44.

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

Lan, S., Palacios, J. A., Karcher, M., Minin, V., and Shahbaba, B. (2015). An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics.

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

Louis, T. A. (1982). Finding the observed information matrix whe nusing the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

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

McVean, G. and Cardin, N. (2005). Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360(1459):1387–1393.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471.

Murray, I., Adams, R. P., and MacKay, D. J. (2010). Elliptical slice sampling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9, pages 541–548.

Palacios, J. A. and Minin, V. N. (2013). Gaussian process-based Bayesian nonparametric inference of population trajectories from gene genealogies. *Biometrics*, 63:8–18.

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., Kivisild, T., Zhai, W., Eriksson, A., Manica, A., Orlando, L., De La Vega, F. M., Tridico, S., Metspalu, E., Nielsen, K., Ávila Arcos, M. C., Moreno-Mayar, J. V., Muller, C., Dortch, J., Gilbert, M. T. P., Lund, O., Wesolowska, A., Karmin, M., Weinert, L. A., Wang, B., Li, J., Tai, S., Xiao, F., Hanihara, T., van Driem, G., Jha, A. R., Ricaut, F.-X., de Knijff, P., Migliano, A. B., Gallego Romero, I., Kristiansen, K., Lambert, D. M., Brunak, S., Forster, P., Brinkmann, B., Nehlich, O., Bunce, M., Richards, M., Gupta, R., Bustamante, C. D., Krogh, A., Foley, R. A., Lahr, M. M., Balloux, F., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Wang, J., and Willerslev, E. (2011). An aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, 334(6052):94–98.

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

Sainudiin, R., Stadler, T., and Véber, A. (2014). Finding the best resolution for the Kingman-Tajima coalescent: theory and applications. *Journal of Mathematical Biology*, pages 1–41.

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

Shahbaba, B., Lan, S., Johnson, W., and Neal, R. (2014). Split Hamiltonian Monte Carlo. *Statistics and Computing*, 24(3):339–349.

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.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.

Wilton, P. R., Carmi, S., and Hobolth, A. (2015). The SMC′ is a highly accurate approximation to the ancestral recombination graph. *Genetics*.

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

Appendix A

Discretization

For both our Bayesian method and our EM method, we assume that $N(t)$ is a piece-wise linear (or piece-wise constant) function with d change points. Let $\mathbf{x}^i = \{x^i_1, x^i_2, \ldots, x^i_{d+n-1}\}$ be the ordered set of time points corresponding to the change points $x_1, \ldots, x_d$ and the coalescent time points $\mathbf{t}^i$ of local genealogy i. Then, we calculate all the factors needed for the observed data likelihood (Equation 13) and the complete data likelihood (Equation 20).

Let $\widetilde{F}^i_{k,j}$ denote the discretized version of F^i that represents the number of branches in g_i that do not coalesce with any other branch in the time interval (x^i_k, x^i_{j+1}). Note that the indices here are in increasing order, $k \leq j$. Similarly, let $\frac{1}{l_i}\widetilde{P}^i_{k,j}$ denote the probability that U_i (the pruning time along genealogy i), occurs in (x^i_k, x^i_{k+1}) and the self-coalescing event occurs at time t^i_{new} in (x^i_j, x^i_{j+1}). That is,

$$\widetilde{P}^i_{k,j} = \begin{cases} \frac{1}{A^i(x^i_{j+1})}(\Delta^i_j - \widetilde{Q}^i_j) & k = j \\ \widetilde{Q}^i_k \tilde{q}^i_{k+1}\tilde{q}^i_{k+2}\cdots\tilde{q}^i_{j+1}(1-\tilde{q}^i_j) & k < j \end{cases}, \tag{28}$$

where

$$\frac{1}{l_i}\widetilde{Q}^i_k = \frac{1}{l_i}\frac{N(x^i_{k+1})}{A^i(x^i_{k+1})}[1-\tilde{q}^i_k] \tag{29}$$

is the joint probability of pruning time $U_i \in (x^i_k, x^i_{k+1})$ and not coalescing back to the same branch in the time interval (x^i_k, x^i_{k+1}), and

$$\tilde{q}^i_k = \exp\left\{-\frac{A^i(x^i_{k+1})\Delta^i_k}{N(x^i_{k+1})}\right\}$$

Expectation-Maximization Algorithm

E-step: Equations 22 and 23 show that for the E-step, the only expectations we need are $\mathrm{E}[z^i_j \mid \mathbf{Y}]$ and $\mathrm{E}[\Delta^i_j \mid \mathbf{Y}]$. We compute these expression as follows:

$$\mathrm{E}[z_j^i \mid \mathbf{Y}] = \begin{cases} \frac{\sum_{k=1}^{j} \widetilde{F}_{k,j}^i \widetilde{P}_{k,j}^i}{\sum_{j=1}^{D} \sum_{k=j}^{D} \widetilde{F}_{k,j}^i \widetilde{P}_{k,j}^i}, & \text{for } i \in \mathcal{I} \text{ (invisible).} \\ z_j^i, & \text{for } i \in \mathcal{I}^c \text{ (visible).} \end{cases}$$

For $i \in \mathcal{I}$

$$\mathrm{E}[\Delta_j^i \mid Y] = (x_{j+1}^i - x_j^i) \frac{\sum_{k=1}^{j-1} \sum_{l=j+1}^{D} \widetilde{F}_{k,l}^i \widetilde{P}_{k,l}^i}{\sum_{j=1}^{D} \sum_{k=j}^{D} \widetilde{F}_{k,j}^i \widetilde{P}_{k,j}^i}$$

$$+ \int_{x_j^i}^{x_{j+1}^i} (x_{j+1}^i - u) \exp\left\{ -\frac{(x_{j+1}^i - u) A^i(x_{j+1}^i)}{N(x_{j+1}^i)} \right\} du \frac{\sum_{k=j+1}^{D} \widetilde{F}_{j,k}^i \frac{\widetilde{P}_{j,k}^i}{\widetilde{Q}_j^i}}{\sum_{j=1}^{D} \sum_{k=j}^{D} \widetilde{F}_{k,j}^i \widetilde{P}_{k,j}^i}$$

$$+ \int_{x_j^i}^{x_{j+1}^i} \int_{u}^{x_{j+1}^i} (t - u) \frac{1}{N(x_{j+1}^i)} \exp\left\{ -\frac{(t - u) A^i(x_{j+1}^i)}{N(x_{j+1}^i)} \right\} dt du \frac{\widetilde{F}_{j,j}^i}{\sum_{j=1}^{D} \sum_{k=j}^{D} \widetilde{F}_{k,j}^i \widetilde{P}_{k,j}^i}$$

$$+ \int_{x_j^i}^{x_{j+1}^i} (t - x_j^i) \exp\left\{ -\frac{(t - x_j^i) A^i(x_{j+1}^i)}{N(x_{j+1}^i)} \right\} dt \frac{\sum_{k=1}^{j-1} \widetilde{F}_{k,j}^i \frac{\widetilde{P}_{k,j}^i}{1-\hat{q}_j^i}}{\sum_{j=1}^{D} \sum_{k=j}^{D} \widetilde{F}_{k,j}^i \widetilde{P}_{k,j}^i},$$

and for $i \in \mathcal{I}^c$, let

$$y_j^i = \begin{cases} 1, & \text{if } t_{new}^i \geq x_{k+1}^i, \\ 0, & \text{othewise.} \end{cases}$$

Then,

$$\mathrm{E}[\Delta_j^i \mid \mathbf{Y}] = \begin{cases} 0, & \text{if } \sum_{k=1}^{j} I^i(x_{k+1}^i) = 0 \text{ or } y_j^i = 0, \\ x_{j+1}^i - x_j^i, & \text{if } I^i(x_{j+1}^i) = 0, \text{ and } \sum_{k=1}^{j} I^i(x_{k+1}^i) > 0, \text{ and } y_j^i = 1, \\ \delta_j^i, & \text{otherwise.} \end{cases}$$

where

$$\delta_j^i = (x_{j+1}^i - x_j^i) \left[\frac{\sum_{k=1}^{j-1} I^i(x_{k+1}^i) \widetilde{Q}_k^i \prod_{l=k+1}^{D_i - 1} \left[\hat{q}_l^i\right]^{y_l^i}}{\sum_{k=1}^{D} I^i(x_{k+1}^i) \widetilde{Q}_k^i \prod_{l=k+1}^{D} \left[\hat{q}_l^i\right]^{y_l^i}} \right] \tag{30}$$

$$+ \int_{x_j^i}^{x_{j+1}^i} (x_{j+1}^i - u) \exp\left\{ -\frac{(x_{j+1}^i - u) A^i(x_{j+1})}{N(x_{j+1}^i)} \right\} du \left[\frac{I^i(x_{j+1}^i) \prod_{l=j+1}^{D} [\hat{q}_l^i]^{y_l^i}}{\sum_{k=1}^{D} I^i(x_{k+1}^i) \widetilde{Q}_k^i \prod_{l=k+1}^{D} \left[\hat{q}_l^i\right]^{y_l^i}} \right]$$

M-step. Now, for the kth iteration of the algorithm and maximizing the complete data log-likelihood (Equation 20) we have

$$N_l^k = \frac{\sum_{j=1}^{D} 0.5 A^0(x_{j+1}^0)[A^0(x_{j+1}^0) - 1](x_{j+1}^0 - x_j^0) 1_{l,j}^0 + \sum_{i=0}^{m-2} \sum_{j=1}^{D} A^i(x_{j+1}^i) \mathrm{E}_{\mathbf{N}^{k-1}}[\Delta_j^i \mid \mathbf{Y}] 1_{l,j}^i}{\sum_{j=1}^{D} a_j^0 1_{l,j}^0 + \sum_{i=0}^{m-2} \sum_{j=1}^{D} \mathrm{E}_{\mathbf{N}^{k-1}}[z_j^i] 1_{l,j}^i}$$

where

$$1_{l,j}^i = \begin{cases} 1, & \text{if } x_l < x_{j+1}^i \leq x_{l+1}, \\ 0, & \text{otherwise.} \end{cases}$$

is an indicator function that takes the value of 1 when (x_l, x_{l+1}) covers the interval (x^i_j, x^i_{j+1}).

Observed score function for Split Hamiltonian Monte Carlo

Our Bayesian approach relies on Split Hamiltonian Monte Carlo (splitHMC) to sample from the posterior distribution of model parameters. This method requires the calculation of the observed score function. We use Fisher's identity and calculate the observed score function as the conditional expected complete score function. The lth element of $\nabla\mathcal{L}_{obs}$ is

$$(\nabla\mathcal{L}_{obs})_l = -\sum_{j=1}^{D-1} a^0_j 1^0_{l,j} + \frac{1}{2}\sum_{j=1}^{D-1} C^0(x^0_{j+1})[C^0(x^0_{j+1}) - 1](x^0_{j+1} - x^0_j)1^0_{l,j}\exp[-\log N_l]$$

$$-\sum_{i=0}^{m-2}\sum_{j=1}^{D-1} E[z^i_j \mid \mathbf{Y}]1^0_{l,j} + \sum_{i=0}^{m-2}\sum_{j=1}^{D-1} C^i(x^i_{j+1})E[\Delta^i_j \mid \mathbf{Y}]1^i_{l,j}\exp[-\log N_l] \quad (31)$$

Fisher Information Calculation

The calculation of the Fisher information needed to estimate confidence intervals of a piece-wise constant trajectory of population sizes, requires the following expected values:

$$\mathrm{E}[z^i_j z^l_k \mid \mathbf{Y}] = \begin{cases} \mathrm{E}[z^i_j \mid \mathbf{Y}] & j = k, i = l \\ 0 & j \neq k, i = l \\ \mathrm{E}[z^i_j \mid \mathbf{Y}]\mathrm{E}[z^l_k \mid \mathbf{Y}] & i \neq l \end{cases}$$

$$\mathrm{E}[\Delta^i_j \Delta^l_k \mid \mathbf{Y}] = \begin{cases} \Delta^i_{j,k} & i = l \\ \mathrm{E}[\Delta^i_j \mid \mathbf{Y}]\mathrm{E}[\Delta^l_k \mid \mathbf{Y}] & i \neq l \end{cases}$$

$$\mathrm{E}[z^i_j \Delta^l_k \mid \mathbf{Y}] = \begin{cases} \mathrm{E}[z^i_j \mid \mathbf{Y}]\mathrm{E}[\Delta^l_k \mid \mathbf{Y}] & i \neq l \\ \mathrm{E}[z^i_j \Delta^i_j \mid \mathbf{Y}] & i = l, j = k \\ 0 & i = l, j < k \\ (z\Delta)^i_{jk} & k < j \end{cases}$$

For $k < j$ and $i \in \mathcal{I}$

$$(z\Delta)^i_{jk} = (x^i_{k+1} - x^i_k)\frac{\sum_{l=1}^{k-1}\hat{F}^i_{l,j}\hat{P}^i_{l,j}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}\hat{P}^i_{k,j}}$$

$$+\int_{x^i_k}^{x^i_{k+1}}(x^i_{k+1} - u)\exp\left\{-\frac{(x^i_{k+1} - u)C^i(x^i_{k+1})}{N(x^i_{k+1})}\right\}du\frac{\hat{F}^i_{k,j}\frac{\hat{P}^i_{k,j}}{\hat{Q}^i_k}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}\hat{P}^i_{k,j}}$$

and for $k < j$ and $i \in \mathcal{I}^c$

$$(z\Delta)^i_{jk} = z^i_j E[\Delta^i_k \mid \mathbf{Y}]$$

For $j < k$ and $i \in \mathcal{I}$

$$\Delta^i_{j,k} = (x^i_{j+1} - x^i_j)(x^i_{k+1} - x^i_k)\frac{\sum_{l=1}^{j-1}\sum_{m=k+1}^{D_i-1}\hat{F}^i_{l,m}\hat{P}^i_{l,m}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}\hat{P}^i_{k,j}}$$

$$+(x^i_{k+1}-x^i_k)\int_{x^i_j}^{x^i_{j+1}}(x^i_{j+1}-u)\exp\left\{-\frac{(x^i_{j+1}-u)C^i(x^i_{j+1})}{N(x^i_{j+1})}\right\}du\frac{\sum_{l=k+1}^{D_i-1}\hat{F}^i_{j,l}\frac{\hat{P}^i_{j,l}}{\hat{Q}^i_j}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}P^i_{k,j}}$$

$$+\int_{x^i_j}^{x^i_{j+1}}\int_{x^i_k}^{x^i_{k+1}}(x^i_{j+1}-u)(t-x^i_k)\exp\left\{-\frac{(x^i_{j+1}-u)C^i(x^i_{j+1})}{N(x^i_{j+1})}-\frac{(t-x^i_k)C^i(x^i_{k+1})}{N(x^i_{k+1})}\right\}dudt\frac{\hat{F}^i_{j,k}\frac{\hat{P}^i_{j,k}}{\hat{Q}^i_j(1-q^i_k)}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}\hat{P}^i_{k,j}}$$

and

$$\Delta^i_{j,j}=(x^i_{j+1}-x^i_j)^2\frac{\sum_{k=1}^{j-1}\sum_{l=j+1}^{D_i-1}\hat{F}^i_{k,l}\hat{P}^i_{k,l}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}\hat{P}^i_{k,j}}$$

$$+\int_{x^i_j}^{x^i_{j+1}}(x^i_{j+1}-u)^2\exp\left\{-\frac{(x^i_{j+1}-u)C^i(x^i_{j+1})}{N(x^i_{j+1})}\right\}du\frac{\sum_{k=j+1}^{D_i-1}\hat{F}^i_{j,k}\frac{\hat{P}^i_{j,k}}{\hat{Q}^i_j}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}P^i_{k,j}}$$

$$+\int_{x^i_j}^{x^i_{j+1}}\int_{u}^{x^i_{j+1}}(t-u)^2\frac{1}{N(x^i_{j+1})}\exp\left\{-\frac{(t-u)C^i(x^i_{j+1})}{N(x^i_{j+1})}\right\}dtdu\frac{\hat{F}_{j,j}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}\hat{P}^i_{k,j}}$$

$$+\int_{x^i_j}^{x^i_{j+1}}(t-x^i_j)^2\exp\left\{-\frac{(t-x^i_j)C^i(x^i_{j+1})}{N(x^i_{j+1})}\right\}dt\frac{\sum_{k=1}^{j-1}\hat{F}^i_{k,j}\frac{\hat{P}^i_{k,j}}{1-\hat{q}^i_j}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}\hat{P}^i_{k,j}}$$

For $i\in\mathcal{I}^c$ and $j<k$

$$\Delta^i_{j,k}=\begin{cases}0 & \sum_{l=1}^{j}I^i(x^i_{l+1})=0\text{ or }y^i_j=0\\(x^i_{j+1}-x^i_j)(x^i_{k+1}-x^i_k) & I^i(x^i_{j+1})=0,\sum_{l=1}^{j}I^i(x^i_{l+1})>0,y^i_j=1,y^i_k=1\\(x^i_{k+1}-x^i_k)\delta^i_j & I^i(x^i_{j+1})=1,\sum_{l=1}^{j}I^i(x^i_{l+1})>0,y^i_k=1\end{cases},$$

where δ^i_j is as defined in Equation 30, and

$$\Delta^i_{j,j}=\begin{cases}0 & \sum_{l=1}^{j}I^i(x^i_{l+1})=0\text{ or }y^i_j=0\\(x^i_{j+1}-x^i_j)^2 & I^i(x^i_{j+1})=0,\sum_{l=1}^{j}I^i(x^i_{l+1})>0,y^i_j=1\\\delta^i_{j,j} & I^i(x^i_{j+1})=1,\sum_{l=1}^{j}I^i(x^i_{l+1})>0,y^i_j>0\end{cases},$$

where

$$\delta^i_{j,j}=(x^i_{j+1}-x^i_j)^2\left[\frac{\sum_{k=1}^{j-1}I^i(x^i_{k+1})\hat{Q}^i_k\prod_{l=k+1}^{D_i-1}\left[\hat{q}^i_l\right]^{y^i_l}}{\sum_{k=1}^{D_i-1}I^i(x^i_{k+1})\hat{Q}^i_k\prod_{l=k+1}^{D_i-1}\left[\hat{q}^i_l\right]^{y^i_l}}\right]$$

$$+\int_{x^i_j}^{x^i_{j+1}}(x^i_{j+1}-u)^2\exp\left\{-\frac{(x^i_{j+1}-u)C^i(x_{j+1})}{N(x^i_{j+1})}\right\}du\left[\frac{I^i(x^i_{j+1})\prod_{l=j+1}^{D_i-1}[\hat{q}^i_l]^{y^i_l}}{\sum_{k=1}^{D_i-1}I^i(x^i_{k+1})\hat{Q}^i_k\prod_{l=k+1}^{D_i-1}\left[\hat{q}^i_l\right]^{y^i_l}}\right]$$

and

For $i\in\mathcal{I}$

$$\mathrm{E}[z^i_j\Delta^i_j\mid\mathbf{Y}]=\int_{x^i_j}^{x^i_{j+1}}\int_{u}^{x^i_{j+1}}(t-u)\frac{1}{N(x^i_{j+1})}\exp\left\{-\frac{(t-u)C^i(x^i_{j+1})}{N(x^i_{j+1})}\right\}dtdu\frac{\hat{F}_{j,j}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}^i_{k,j}\hat{P}^i_{k,j}}$$

$$+\int_{x_j^i}^{x_{j+1}^i}(t-x_j^i)\exp\left\{-\frac{(t-x_j^i)C^i(x_{j+1}^i)}{N(x_{j+1}^i)}\right\}dt\frac{\sum_{k=1}^{j-1}\hat{F}_{k,j}^i\frac{\hat{P}_{k,j}^i}{1-\hat{q}_j^i}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}_{k,j}^i\hat{P}_{k,j}^i}.$$

and for $i \in \mathcal{I}^c$

$$\mathrm{E}[z_j^i\Delta_j^i \mid \mathbf{Y}] = \delta_j^i z_j^i$$

The gradient vector of the complete data log-likelihood has lth element

$$\frac{\partial}{\partial \log N_l}\mathcal{L}_c(\mathbf{Y}_c;\hat{\mathbf{N}}) = B_l - A_l + C_l - Z_l \tag{32}$$

With

$$A_l = \sum_{j=1}^{D} a_j^0 1_{l,j}^0$$

$$B_l = \sum_{j=1}^{D} 0.5C^0(x_{j+1}^0)[C^0(x_{j+1}^0)-1](x_{j+1}^0 - x_j^0)1_{l,j}^0\exp[-\log N_l],$$

$$C_l = \sum_{i=0}^{m-2}\sum_{j=1}^{D} C^i(x_{j+1}^i)\Delta_j^i 1_{l,j}^i\exp[-\log N_l],$$

and

$$Z_l = \sum_{i=0}^{m-2}\sum_{j=1}^{D} z_j^i 1_{l,j}^i.$$

Next, differentiating Equation (32), we have $\frac{\partial^2 l_c(\mathbf{Y}_c;\hat{\mathbf{N}})}{\partial \log N_l \partial \log N_m} = 0$ for all $l \neq m$, so the Hessian is a diagonal matrix with (l,l)th element

$$\frac{\partial^2}{\partial \log N_l^2}\mathcal{L}_c(\mathbf{Y}_c;\hat{\mathbf{N}}) = -B_l - C_l$$

and

$$\mathrm{E}\left[\left(\frac{\partial l_c(\mathbf{Y}_c;\hat{\mathbf{N}})}{N_l}\right)^2 \mid \mathbf{Y}\right] = (B_l - A_l)^2 + 2(B_l - A_l)\mathrm{E}[C_l - Z_l \mid \mathbf{Y}] + \mathrm{E}[(C_l - Z_l)^2 \mid \mathbf{Y}]$$

where

$$\mathrm{E}[C_l^2 \mid \mathbf{Y}] = \exp[-2logN_l]\sum_{i=0}^{m-2}\sum_{j=1}^{D_i-1}\left[\left\{C^i(x_{j+1}^i)\right\}^2\Delta_{j,j}^i 1_{l,j}^i + 2\sum_{k=j+1}^{D_i-1}C^i(x_{j+1}^i)C^i(x_{k+1}^i)\Delta_{j,k}^i 1_{l,j}^i 1_{l,k}^i\right]$$

$$+2\exp[-2logN_l]\sum_{i=0}^{m-2}\sum_{j=1}^{D_i-1}\left[C^i(x_{j+1}^i)\mathrm{E}[\Delta_j^i \mid \mathbf{Y}]1_{l,j}^i\sum_{p=i+1}^{m-2}\sum_{k=1}^{D_p-1}C^p(x_{k+1}^p)\mathrm{E}[\Delta_k^p \mid \mathbf{Y}]1_{l,k}^p\right],$$

$$\mathrm{E}[Z_l^2 \mid \mathbf{Y}] = \sum_{i=0}^{m-2}\sum_{j=1}^{D_i-1}\left[\mathrm{E}[z_j^i \mid \mathbf{Y}]1_{l,j}^i + 2\mathrm{E}[z_j^i \mid \mathbf{Y}]1_{l,j}^i\sum_{p=i+1}^{m-2}\sum_{k=1}^{D_p-1}\mathrm{E}[z_k^p \mid \mathbf{Y}]1_{l,k}^p\right]$$

and

$$\mathrm{E}[C_l Z_l \mid \mathbf{Y}] = \frac{1}{N_l}\sum_{i=0}^{m-2}\sum_{j=1}^{D} C^i(x^i_{j+1})\mathrm{E}[z^i_j\Delta^i_j \mid \mathbf{Y}]1^i_{l,j} + \sum_{i=0}^{m-2}\sum_{j=1}^{D} C^i(x^i_{j+1})\sum_{k=j+1}^{D}(z\Delta)^i_{k,j}1^i_{l,k}1^i_{l,j}$$

$$+\sum_{i=0}^{m-2}\sum_{j=1}^{D} C^i(x^i_{j+1})\sum_{p=1,p\neq i}^{m-2}\sum_{k=1}^{D_p-1}\mathrm{E}[\Delta^i_j \mid \mathbf{Y}]\mathrm{E}[z^p_k \mid \mathbf{Y}]1^i_{l,j}1^p_{l,k}$$

Also,

$$\mathrm{E}\left[\left(\frac{\partial l_c(\mathbf{Y}_c;\hat{\mathbf{N}})}{N_l}\right)\left(\frac{\partial l_c(\mathbf{Y}_c;\hat{\mathbf{N}})}{N_k}\right) \mid \mathbf{Y}\right] = (B_l - A_l)(B_k - A_k) + (B_l - A_l)\mathrm{E}[C_k - Z_k \mid \mathbf{Y}]$$

$$+(B_k - A_k)\mathrm{E}[C_l - Z_l \mid \mathbf{Y}] + \mathrm{E}[(C_l - Z_l)(C_k - Z_k) \mid \mathbf{Y}]$$

where

$$\mathrm{E}[C_l C_k] = \exp[-logN_l - logN_k]\sum_{i=0}^{m-2}\sum_{j=1}^{D_i-1} C^i(x^i_{j+1})1^i_{l,j}\sum_{o=1}^{m-2}\sum_{p=1}^{D_o-1} C^o(x^o_{p+1})1^o_{k,p}\mathrm{E}[\Delta^i_j\Delta^o_p \mid \mathbf{Y}],$$

$$\mathrm{E}[Z_l Z_k] = \sum_{i=0}^{m-2}\sum_{j=1}^{D_i-1}\sum_{o\neq i}^{m-2}\sum_{p=1}^{D_o-1} 1^i_{l,j}1^o_{k,p}\mathrm{E}[z^i_j z^o_p \mid \mathbf{Y}]$$

and for $l < o$

$$\mathrm{E}[C_l Z_o \mid \mathbf{Y}] = \frac{1}{N_l}\sum_{i=0}^{m-2}\sum_{j=1}^{D} C^i(x^i_{j+1})1^i_{l,j}\left\{\sum_{k=j+1}^{D}(z\Delta)^i_{k,j}1^i_{o,k} + \sum_{p=1,p\neq i}^{m-2}\sum_{k=1}^{D}\mathrm{E}[\Delta^i_j \mid \mathbf{Y}]\mathrm{E}[z^p_k \mid \mathbf{Y}]1^p_{o,k}\right\}$$

Sufficient statistics under SMC′

Proposition 1. *For a single locus, the set of coalescent times are sufficient statistics for inferring $N(t)$.*

Proof. This can be proved using the factorization theorem. The marginal density of a local genealogy (Equation 3) has a unique factor that depends on $N(t)$ and g only through $t_n, \ldots, t_2$. The values of $A(t)$ are induced by the natural order of the coalescent times. ■

Let F denote a lower triangular matrix of size $n \times n$ with the $F_{i,j}$ entry the number of lineages that do not coalesce in the time interval (t_{i+1}, t_j), as defined in the methods section and with the following properties:

1. $F_{i,1} = 0$ for all $i = 1, \dots, n$ (The first column contains 0s for completion)

2. $F_{i,j} = 0$ for all $j > i$ (Lower triangular matrix)

3. $F_{i,i} = i$ for all $i \geq 2$ (The diagonal corresponds to the number of lineages at each intercoalescent interval)

4. $F_{i,i-1} = i - 2$ for all $i \geq 2$ (At each intercoalescent interval, we loose two free lineages, so the second diagonal correspond to the number of lineages minus two)

5. For $j < n - 1$, the last row of F is defined according to:
$$F_{n,j-1} = \begin{cases} F_{n,j} - 2, & \text{with probability } p = \frac{\binom{F_{n,j}}{2}}{\binom{j}{2}}, \\ F_{n,j} - 1, & \text{with probability } p = \frac{F_{n,j}(j - F_{n,j})}{\binom{j}{2}}, \\ F_{n,j}, & \text{with probability } p = \frac{\binom{j - F_{n,j}}{2}}{\binom{j}{2}} \end{cases}$$

6. Let c denote the number of cherries, then
$$c = \sum_{j=2}^{n} 1_{\{F_{n,j} - F_{n,j-1} = 2\}}$$

7. For $i < n$ and $j < i - 1$, if $F_{n,j-1} = F_{n,j} - 2$, then $F_{i,j-1} = F_{i,j} - 2$.

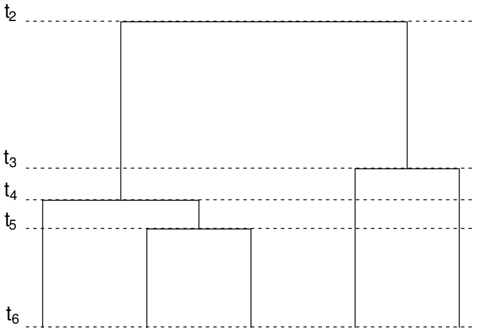
8. Let v_i denote the set of lineages in the intercoalescent interval (t_i, t_{i-1}) with direct descendant internal nodes. The lineage labels correspond to the label of the coalescent time, when the direct descendant internal node was created. That is, the lineage created at t_n has label n: $v_n = \{n\}$; the lineage created at t_i has label i. Let $|v_i|$ denote the size of the set v_i. Note that $1 \leq |v_i| \leq c$ and
$$|v_i| = \sum_{j=i}^{n} 1_{\{F_{n,j} - F_{n,j-1} = 2\}} - \sum_{j=i}^{n} 1_{\{F_{n,j} - F_{n,j-1} = 0\}}$$

9. For $i < n$ and $j < i - 1$, if $F_{n,j-1} = F_{n,j} - 1$, then at time t_j, there is a coalescence between a singleton and a lineage in the set v_j. Let a_j be the lineage selected uniformly at random from v_j, then
$$F_{i,j-1} = \begin{cases} F_{i,j} - 1 & \text{if } i > a_j \\ F_{i,j} - 2 & \text{if } j < i \leq a_j \end{cases}$$

10. For $i < n$ and $j < i - 1$, if $F_{n,j-1} = F_{n,j}$, then at time t_j, there is a coalescence between two lineages a_j^1 and a_j^2 from the set v_j. Let a_j^1 denote the minimum and a_j^2 the maximum of the two lineages selected, then

$$F_{i,j-1} = \begin{cases} F_{i,j} & \text{if } i > a_j^2 \\ F_{i,j} - 1 & \text{if } a_j^1 < i \leq a_j^2 \\ F_{i,j} - 2 & \text{if } j < i \leq a_j^1 \end{cases}$$

We show the correspondence between a ranked tree shape and the F-matrix in the example of Figure A1. The first row and the first column are set to 0, the first two diagonals are known with probability 1: $F_{i,i} = i$ and $F_{i,i-1} = F_{i,i} - 2$ for $i > 1$. In our example, $n = 5$ and so, the first diagonal corresponds to $(0, 2, 3, 4, 5)$ and the second diagonal corresponds to $(0, 1, 2, 3)$. The last row $F_{5,}$ contains 0, followed by the number of branches that do not coalesce in the time intervals (t_6, t_2), (t_6, t_3), (t_6, t_4) and (t_6, t_5) corresponding to $(0, 0, 2, 3, 5)$.



$$F = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 2 & 4 & 0 \\ 0 & 0 & 2 & 3 & 5 \end{pmatrix}$$

Figure A1: Ranked tree shape for $n = 6$

Proposition 2. *There is a bijection between the set of ranked tree shapes $\mathcal{H}_n$ and $\mathcal{F}$, the set of F-matrices.*

Proof. The probability of the F matrix can be expressed as the product of the conditional probabilities of the columns of the F matrix, that is:

$$\Pr(F) = \Pr(F_{\cdot,n}) \prod_{j=1}^{n-1} \Pr(F_{\cdot,n-j} \mid F_{\cdot,n-j+1})$$
$$= \prod_{j=2}^{n-2} \Pr(F_{\cdot,n-j} \mid F_{\cdot,n-j+1}),$$

since the first and last column of F are known with probability 1. Note $F_{\cdot,j}$ represents the j-th column vector of the F matrix.

Let $d_i = F_{n,i} - F_{n,i-1}$ for $i = 3, \ldots, n$, and $d_2 = F_{n,2}$ then

$$\Pr(F_{\cdot,n-j} \mid F_{\cdot,n-j+1}) = \Pr(d_{n-j} \mid F_{\cdot,n-j+1})\Pr(F_{n-j:n-1,n-j} \mid d_{n-j}, F_{\cdot,n-j+1}). \tag{33}$$

That is, the conditional probability of the $(n-j)$th column of F given the $(n-j+1)$th column of F is the product of the conditional probability of the last element of the $(n-j)$th column and the conditional probability of the rest of the $(n-j)$th column. When $d_{n-j} = 2$ the rest of the column is known with probability 1 (property 7 of the F-matrix). When $d_{n-j} = 1$, the rest of the $n-j$th column has probability $1/|v_{n-j+1}|$ (property 9 of the $F-$matrix) and when $d_{n-j} = 2$, the rest of the $n-j$th column has probability $1/\binom{|v_{n-j+1}|}{2}$ (property 10 of the F-matrix). Then re-writing Equation 33, we have

$$\Pr(F_{\cdot,n-j} \mid F_{\cdot,n-j+1}) = \Pr(d_{n-j+1} \mid F_{\cdot,n-j+1}) \left(\frac{1}{|v_{n-j+1}|}\right)^{1_{\{d_{n-j}=1\}}} \left(\frac{1}{\binom{|v_{n-j+1}|}{2}}\right)^{1_{\{d_{n-j}=0\}}}, \tag{34}$$

since $|v_{n-j}| = \sum_{k=n-j}^{n}(1_{\{d_k=2\}} - 1_{\{d_k=0\}})$, and $F_{n,k} = n - \sum_{j=k+1}^{n} d_j$, then

$$\Pr(d_{n-j+1} \mid F_{\cdot,n-j+1}) = \Pr(d_{n-j+1} \mid F_{n,n-j+1}) = \Pr(d_{n-j+1} \mid \sum_{k=n-j+2}^{n} d_k),$$

and

$$\begin{aligned}
\Pr(F) &= \prod_{j=1}^{n-2} \Pr(d_{n-j} \mid \sum_{k=n-j+1}^{n} d_k) \left(\frac{1}{|v_{n-j+1}|}\right)^{1_{\{d_{n-j}=1\}}} \left(\frac{2}{|v_{n-j+1}|(|v_{n-j+1}|-1)}\right)^{1_{\{d_{n-j}=0\}}} \\
&= \prod_{j=1}^{n-2} \left(\frac{(n-\sum_{k=n-j+1}^{n} d_k)(n-\sum_{k=n-j+1}^{n} d_k - 1)}{2}\right)^{1_{\{d_{n-j}=2\}}} \\
&\times \left(\frac{(n-\sum_{k=n-j+1}^{n} d_k)(\sum_{k=n-j+1}^{n} d_k - j)}{|v_{n-j+1}|}\right)^{1_{\{d_{n-j}=1\}}} \\
&\times \left(\frac{(\sum_{k=n-j+1}^{n} d_k - j)(\sum_{k=n-j+1}^{n} d_k - j - 1)}{|v_{n-j+1}|(|v_{n-j+1}|-1)}\right)^{1_{\{d_{n-j}=0\}}} \frac{1}{\binom{n-j}{2}} \\
&= \frac{2^{n-2}2^{1-c}}{(n-1)!(n-2)!} \prod_{j=2}^{n-1} \left((n-\sum_{k=j+1}^{n} d_k)(n-\sum_{k=j+1}^{n} d_k - 1)\right)^{1_{\{d_j=2\}}} \\
&\times \left((n-\sum_{k=j+1}^{n} d_k)(j-n+\sum_{k=j+1}^{n} d_k)\right)^{1_{\{d_j=1\}}} \left((j-n+\sum_{k=j+1}^{n} d_k)(j-n+\sum_{k=j+1}^{n} d_k - 1)\right)^{1_{\{d_j=0\}}} \\
&\times \left(\frac{1}{|v_{j+1}|}\right)^{1_{\{d_j=0\}}+1_{\{d_j=1\}}} \left(\frac{1}{|v_{j+1}|-1}\right)^{1_{\{d_j=0\}}}
\end{aligned}$$

Since $d_n = 2$ and $|v_n| = 1$, for $j = n-1$, then d_{n-1} is either 1 or 2, then

$$\begin{aligned}
\Pr(F) &= \frac{2^{n-c-1}}{(n-1)!(n-2)!}(n-2)(n-3)^{1_{\{d_{n-1}=2\}}} \prod_{j=2}^{n-2} \left((n-\sum_{k=j+1}^{n} d_k)(n-\sum_{k=j+1}^{n} d_k - 1)\right)^{1_{\{d_j=2\}}} \\
&\times \left((n-\sum_{k=j+1}^{n} d_k)(j-n+\sum_{k=j+1}^{n} d_k)\right)^{1_{\{d_j=1\}}} \left((j-n+\sum_{k=j+1}^{n} d_k)(j-n+\sum_{k=j+1}^{n} d_k - 1)\right)^{1_{\{d_j=0\}}} \\
&\times \left(\frac{1}{|v_{j+1}|}\right)^{1_{\{d_j=0\}}+1_{\{d_j=1\}}} \left(\frac{1}{|v_{j+1}|-1}\right)^{1_{\{d_j=0\}}}
\end{aligned}$$

If we continue expanding the expressions, we get:

$$\begin{aligned}
\Pr(F) = {} & \frac{2^{n-c-1}}{(n-1)!(n-2)!}(n-2)(n-3)(n-4)^{1_{\{d_{n-2}=2\}}+1_{\{d_{n-2}=1\}}1_{\{d_{n-1}=2\}}}(n-5)^{1_{\{d_{n-2}=2\}}1_{\{d_{n-1}=2\}}} \\
& \times \prod_{j=2}^{n-3}\left((n-\sum_{k=j+1}^{n} d_k)(n-\sum_{k=j+1}^{n} d_k - 1)\right)^{1_{\{d_j=2\}}} \\
& \times \left((n-\sum_{k=j+1}^{n} d_k)(j-n+\sum_{k=j+1}^{n} d_k)\right)^{1_{\{d_j=1\}}} \left((j-n+\sum_{k=j+1}^{n} d_k)(j-n+\sum_{k=j+1}^{n} d_k - 1)\right)^{1_{\{d_j=0\}}} \\
& \times \left(\frac{1}{|v_{j+1}|}\right)^{1_{\{d_j=0\}}+1_{\{d_j=1\}}} \left(\frac{1}{|v_{j+1}|-1}\right)^{1_{\{d_j=0\}}} \\
= {} & \cdots \\
= {} & \frac{2^{n-c-1}}{(n-1)!}
\end{aligned}$$

■

Note that the entries of the F matrix correspond to the same quantity needed to express the transition density of an invisible event (Equation 11). We claim that the sequence of coalescent times sets $\mathbf{t}^0, \mathbf{t}^1, \dots, \mathbf{t}^{m-1}$ and $F^0, F^1, \dots, F^{m-1}$ matrices corresponding to the ranked tree shapes of local genealogies $g_0, g_1, \dots, g_{m-1}$ are sufficient statistics to infer $N(t)$ under the SMC$'$ process. We prove this through the following propositions.

Proposition 3. *The probability density of Tajima's genealogy is proportional, up to a combinatorial factor, to the probability density of Kingman's genealogy.*

Proof.

$$\begin{aligned}
\Pr[G^T = \{F, t_n, t_{n-1}, \dots, t_2\} \mid N(t)] &= \Pr[t_n, t_{n-1}, \dots, t_2 \mid N(t)]\Pr[F \mid t_n, t_{n-1}, \dots, t_2] \\
&= \frac{n!(n-1)!}{2^{n-1}}\Pr[G = \{K_n, t_n, \dots, t_2\} \mid N(t)]\frac{2^{n-c-1}}{(n-1)!} \\
&= \frac{n!}{2^c}\prod_{j=2}^{n}\frac{1}{N(t_j^0)}\exp\left\{-\int_{t_{j+1}^0}^{t_j^0}\frac{A^0(t)(A^0(t)-1)dt}{2N(t)}\right\} \qquad (35)
\end{aligned}$$

■

Proposition 4. *The marginal visible transition density from a local Kingman's genealogy g_{i-1} to G_i is proportional to the marginal visible transition density from the corresponding local Tajima's genealogy g_{i-1}^T to G_i^T.*

Proof. When the labeled topology of g_{i-1} is the same as the labeled topology of g_i, then a transition from g_{i-1} to g_i contains the same information about pruning location as a transition from g^T_{i-1} to g^T_i (Supplementary Information, Figures S1A and S2D). In fact, the $I^{i-1}(t)$ function defined in section 2.1.2 (Equation 8) can be defined in terms of the F^i-matrix and the coalescent times $\mathbf{t}^{i-1}$ and $\mathbf{t}^i$. In this case, for some $j \in \{2, \dots, n\}$, $t^{i-1}_j = t^i_{del}$ and $t^i_j = t^i_{new}$. Then

$$I^{i-1}(t) = \begin{cases} 0, & \text{if } t > \min(t^i_{new}, t^i_{del}), \\ \\ F^{i-1}_{l,j} - F^{i-1}_{l,j-1}, & \text{if } t \in (t^{i-1}_{l+1}, t^{i-1}_l) \text{ for } l = j, j+1, \dots, n. \end{cases}$$

Hence, if $K_{i-1} = K_i$, the labeled topologies of g_{i-1} and g_i, then

$$\Pr[G_i = \{K^{i-1}, \mathbf{t}^i\} \mid g_{i-1} = \{K^{i-1}, \mathbf{t}^{i-1}\}, N(t)] = \Pr[G^T_i = \{F^{i-1}, \mathbf{t}^i\} \mid g^T_{i-1} = \{F^{i-1}, \mathbf{t}^{i-1}\}, N(t)].$$

When the labeled topologies of g_{i-1} and g_i are different, but the children of t^i_{del} and the children of t^i_{new} are the same, we cannot exactly identify the pruning branch and the new coalescing branch (Supplementary Information, Figure S1B) and then a transition from g_{i-1} to g_i contains the same information about pruning location as a transition from g^T_{i-1} to g^T_i. Let $t^{i-1}_j = t^i_{del}$ and $t^i_k = t^i_{new}$, since the children of t^{i-1}_j and t^i_k are the same, it is enough to consider F^{i-1}. Then

$$I^{i-1}(t) = \begin{cases} 0, & \text{if } t > \min(t^i_{new}, t^i_{del}), \\ \\ F^{i-1}_{l,j} - F^{i-1}_{l,j-1}, & \text{if } t \in (t^{i-1}_{l+1}, t^{i-1}_l) \text{ for } l = j, j+1, \dots, n. \end{cases}$$

and

$$\Pr[G_i = \{K^i, \mathbf{t}^i\} \mid g_{i-1} = \{K^{i-1}, \mathbf{t}^{i-1}\}, N(t)] = \Pr[G^T_i = \{F^{i-1}, \mathbf{t}^i\} \mid g^T_{i-1} = \{F^{i-1}, \mathbf{t}^{i-1}\}, N(t)].$$

Now, when the deleted node corresponding to t_{del} is a cherry and the new node corresponding to t_{new} is also a cherry, there are four possible topologies K_i that lead to the same ranked tree shape F^i, then

$$\Pr[G_i = g_i \mid g_{i-1}, N(t)] = \left(\frac{1}{2}\right)^{1_{\{t_j^{i-1}=t_{del}^i\}} 1_{\{F_{n,j}^{i-1}=F_{n,j+1}^{i-1}-2\}}} \times \left(\frac{1}{2}\right)^{1_{\{t_j^i=t_{new}^i\}} 1_{\{F_{n,j}^i=F_{n,j+1}^i-2\}}} \times$$
$$\times \Pr[G_i^T = g_i^T \mid g_{i-1}^T, N(t)],$$

■

Proposition 5. *The marginal invisible transition density from a local Kingman's genealogy g_{i-1} to G_i is equal to the marginal invisible transition density from the corresponding local Tajima's genealogy g_{i-1}^T to G_i^T.*

Proof.

$$\Pr[G_i = g_{i_1} \mid g_{i-1}, N(t)] = \Pr[G_i = g_{i-1} \mid g_{i-1}^T, N(t)],$$

since all needed to compute the transition probability are the coalescent times and the F^{i-1} matrix. Since the topology does not change, the proof follows.

■

Proposition 6. *The Likelihood of partially observed embedded SMC$\,'$ chain of local Kingman's genealogies is proportional, up to a combinatorial factor, to the likelihood of partially observed embedded SMC$\,'$ chain of the corresponding local Tajima's genealogies.*

Proof. The proof follows from propositions 3, 4 and 5 needed to express the likelihood of partially observed embedded SMC$'$ chain (Equation 13).

■

Supporting Information: **Bayesian Nonparametric Inference of Population Size Changes from Sequential Genealogies**

Julia A. Palacios[1,2,3], John Wakeley[1], and Sohini Ramachandran[2,3]

[1]Department of Organismic and Evolutionary Biology, Harvard University
[2]Department of Ecology and Evolutionary Biology, Brown University
[3]Center for Computational Molecular Biology, Brown University

1 Visible Transitions

Figure S1A shows an example of a visible transition when the topology remains the same and Figure S1B shows an example of a visible transition when the topology changes. Green lines mark the possible pruning locations that could have lead to the same visible transition; the red circle indicates the deleted node at coalescent time t_{del} and the blue circle indicates the new node created at coalescent time t_{new}.

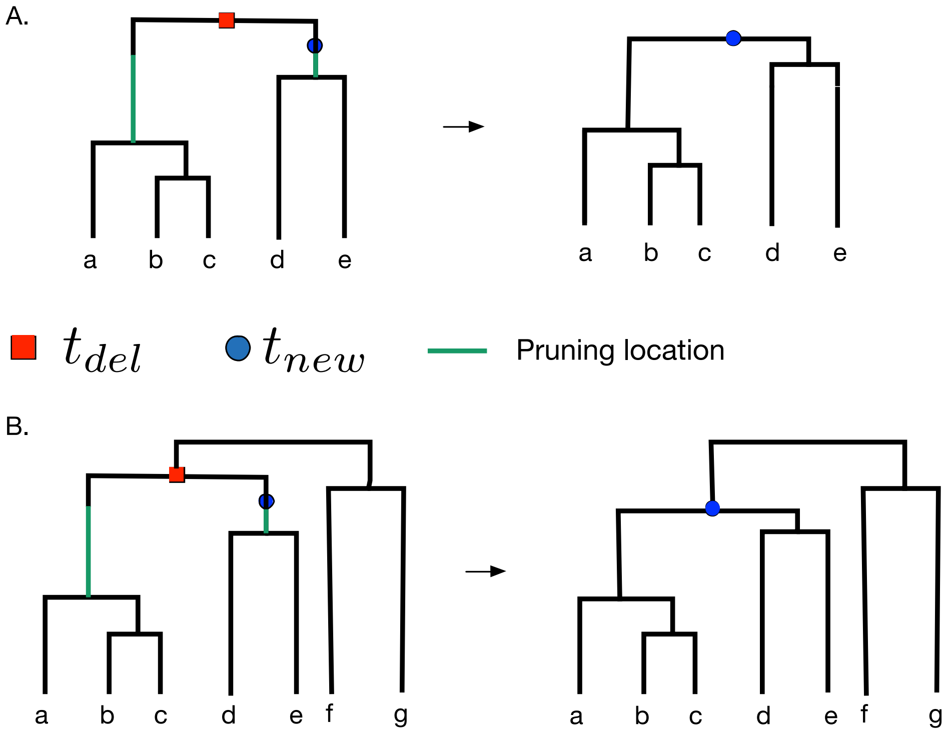


Figure S1: Examples of visible transitions when the pruning branch is uncertain. Red circle indicates deleted node at coalescent time t_{del}, blue circle indicates new node at coalescent time t_{new}. Green lines indicates possible pruning locations that could have resulted in such a visible transition. A. The topology remains the same. B. The topology changes.

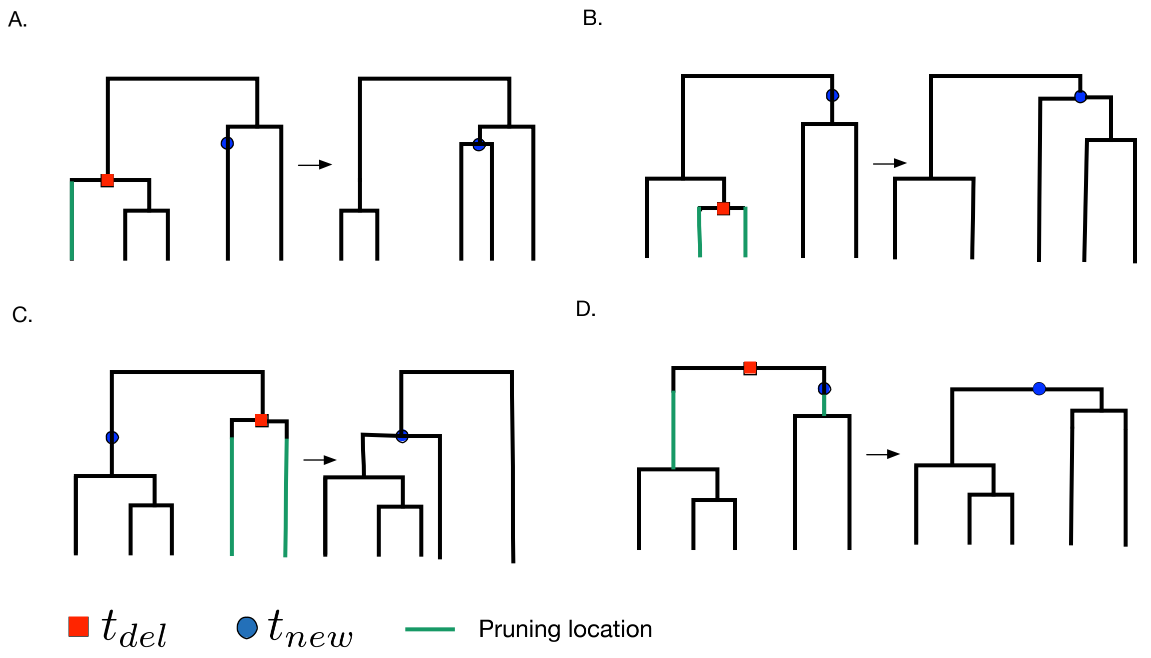


Figure S2: Examples of visible transitions between local Tajima's genealogies. Red circle indicates deleted node at coalescent time t_{del}, blue circle indicates new node at coalescent time t_{new}. Green lines indicates possible pruning locations that could have resulted in such a visible transition.

2 Visible transitions between Tajima's genealogies

A Tajima's genealogy g^T is an unlabeled genealogy. In Figure S2, we show four possible visible transitions. In the first case (Figure 2A), when we compare the number of *children* of the blue circle node on the right tree at time t with the *children* of the red circle node on the left tree, we can conclude that only the green branch could have been selected for pruning. In Figure 2B, comparing the *children* of the blue circle node on the right genealogy to the *children* of the red circle in the left genealogy, we conclude that the two *children* of the red circle are possible pruning locations. In Figures 2C-D, $t_{new} < t_{del}$. This implies that the possible pruning locations will necessarily have heights up to t_{new}. Again, by comparing the *children* of the blue circle node on the right to the *children* of the red circle node on the left, we can asses the possible pruning locations.

3 Simulations with MaCS

We use MaCS (Chen et al., 2009) for all our simulations with the following code lines:
Constant population size:

```
./macs2 300000 -t 1.0 -T -r .005 -h 1 (SEED: 1420480396)
./macs20 3000000 -t 1.0 -T -r .0002 -h 1 (SEED: 1399175725)
./macs100 3000000 -t 1.0 -T -r .0002 -h 1 (SEED: 1400528079)
```

Exponential growth and constant:

```
./macs2 300000 -t 1.0 -eG .1 10 -T -r .02 -h 1 (SEED: 1419985269)
./macs20 300000 -t 4.0 -eG .1 10 -T -r .002 -h 1 (SEED: 1420040333)
./macs100 300000 -t 1.0 -eG .1 10 -T -r .0002 -h 1 (SEED: 1401855826)
```

Bottleneck:

```
./macs2 300000 -t 4.0 -eN 0 1 -eN 0.3 0.1 -eN 0.5 1 -T -r .01 -h 1 (SEED: 1420824821)
./macs20 300000 -t 4.0 -eN 0 1 -eN 0.3 0.1 -eN 0.5 1 -T -r .002 -h 1 (SEED: 1420826310)
./macs100 300000 -t 4.0 -eN 0 1 -eN 0.3 0.1 -eN 0.5 1 -T -r .001 -h 1 (SEED: 1420826409)
```

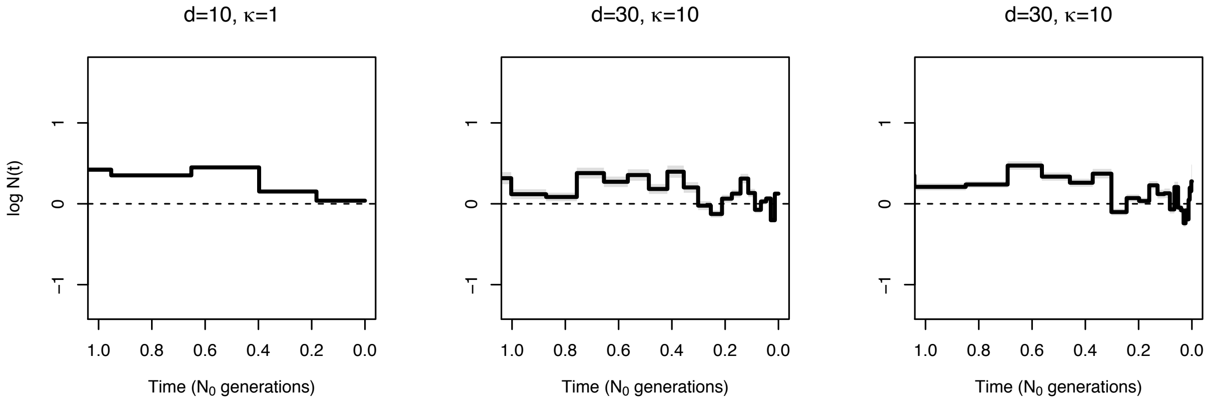


Figure S3: **EM sensitivity to parameter discretization.** Comparison of population size trajectories estimated from 1000 simulated genealogy ($m = 1000$) of 100 individuals with a constant population size. EM inference with different discretizations varying the parameters in Equation **??**.

Table S1: *Summary of simulation results depicted in Figure S3. SRE is the sum of relative errors (Equation 24), MRW is the mean relative width of the 95% BCI (Equation 25), and ENV (Equation 26).*

	SRE	MRW	ENV
EM $d = 10$, $\kappa = 10$	43.41	0.99	48.6%
EM $d = 30$, $\kappa = 10$	34.25	0.76	42.6%
EM $d = 30$, $\kappa = 100$	43.96	0.99	46.0%

4 EM sensitivity to parameter discretization

In Figure S3, we show EM estimates of a constant population size from 1000 local genealogies of 100 individuals. We show that different discretizations result in different estimates. We note that confidence intervals perform poorly in terms of coverage. The performance statistics corresponding to the three estimations displayed in Figure S3 are shown in Table S1.

5 Analysis of Human data

We use *ARGweaver* (Rasmussen et al., 2014) with the following code lines:
European population:

```
arg-sample -s data1000/CEU_10.sites
   -N 11534 -r 1.6e-8 -m 1.26e-8
   --ntimes 200 --maxtime 200e3 -c 1 -n 10
   -o data1000/CEU.sample/out
```

Yoruban population:

```
arg-sample -s data1000/YRI_10.sites
   -N 11534 -r 1.6e-8 -m 1.26e-8
   --ntimes 200 --maxtime 200e3 -c 1 -n 10
   -o data1000/YRI.sample/out
```

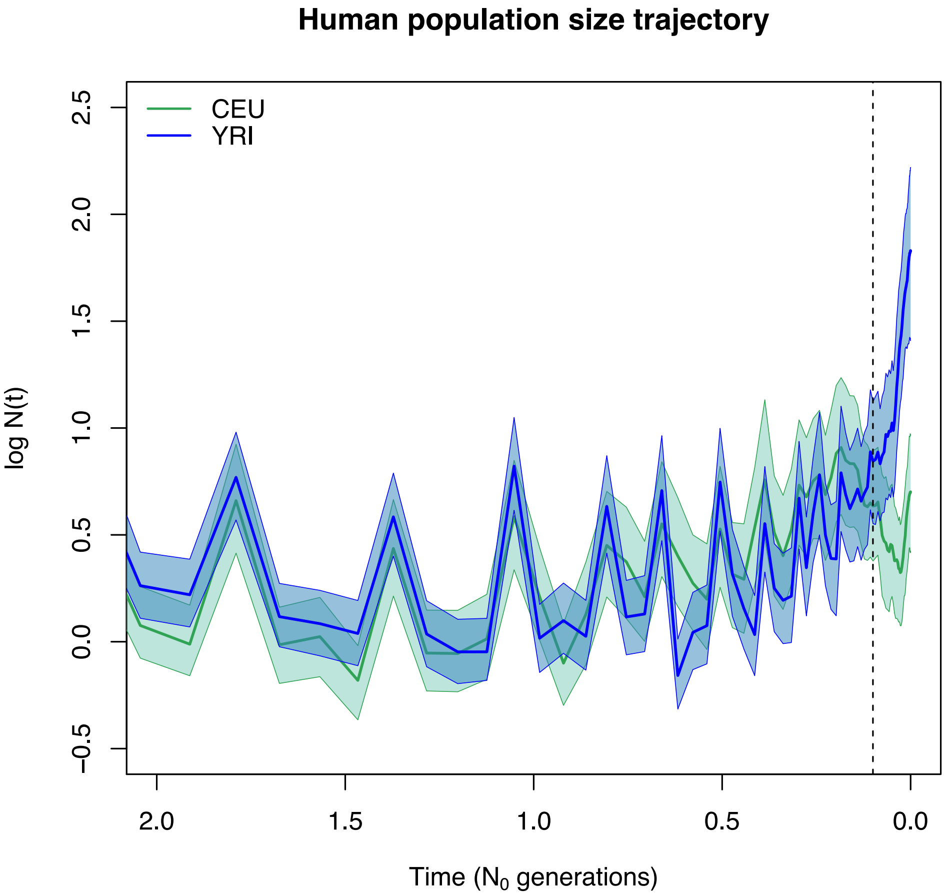


Figure S4: **Inference of human population size trajectories $N(t)$ for $n = 10$.** Green solid line and green shaded areas represent the posterior median and 95% BCI for European population (CEU) and blue solid line and blue shaded areas represent the posterior median and 95% BCI for Yoruban population.

ARGweaver time is measured in units of generations, so in order to generate Figure 8, we multiplied time by $1/(2 \times 11,534)$. To obtain $\log N(t)$ displayed in Figure 8, we multiplied our estimates by $1/(8 \times 11,532)$ and converted them in logarithmic scale.

References

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

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