Bayesian Nonparametric Inference of Population Size Changes from Sequential Genealogies

Bayesian Nonparametric Inference of Population Size Changes from Sequential Genealogies

Julia A. Palacios,*,†,‡,[1] John Wakeley,* and Sohini Ramachandran†,‡,[1]
*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, and †Department of Ecology and Evolutionary Biology and ‡Center for Computational Molecular Biology, Brown University, Providence, Rhode Island 02912

ABSTRACT Sophisticated inferential tools coupled with the coalescent model have recently emerged for estimating past population sizes from genomic data. 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 that 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% of true values across multiple demographic scenarios, enabling formal hypothesis testing about population size differences over time. Using genealogies estimated with *ARG*weaver, 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

FOR a single nonrecombining locus, neutral coalescent theory predicts the set of timed ancestral relationships among sampled individuals, known as a gene genealogy (Kingman 1982; Hudson 1983, 1990; Tajima 1983). In the coalescent model with variable population size, the rate at which two lineages have a common ancestor (or coalesce) 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 nonrecombining locus. 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 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 the 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 SMC′ models, population size trajectories and sequence data are separated by two

doi: 10.1534/genetics.115.177980
Manuscript received May 7, 2015; accepted for publication July 21, 2015; published Early Online July 28, 2015.
Available freely online through the author-supported open access option.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.177980/-/DC1.
[1]Corresponding authors: 80 Waterman St., Box G-W, Brown University, Providence, RI 02912. E-mail: julia.pal.r@gmail.com; E-mail: sramachandran@brown.edu

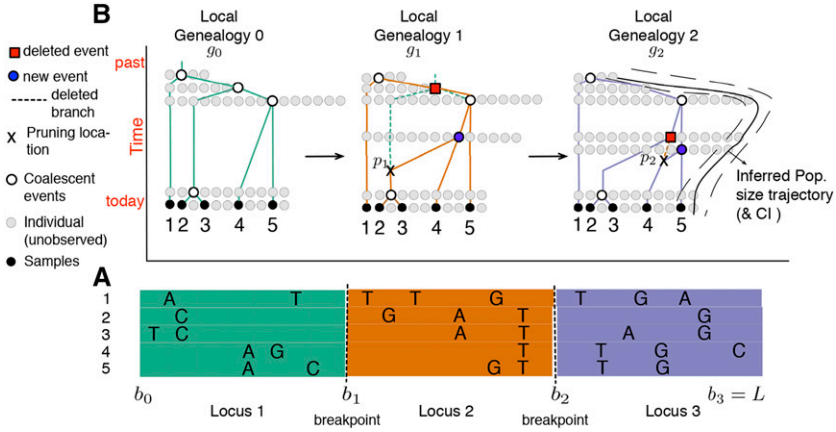


Figure 1 SMC′ hidden Markov model for inferring population size trajectories, drawn according to Rasmussen *et al.* (2014) to highlight notation specific to our study. (A) Observed sequence data in a segment of length L from five individuals. Three loci are shown delimited by recombination breakpoints 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 solid black circles. Local genealogies have a Markovian degree-1 dependency. Each intercoalescent time (the time interval between coalescent events denoted as open circles) provides information about past population size (number of solid gray 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 (solid blue circle). The coalescent event of g_0 depicted as a solid red 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.

stochastic processes: (i) *a state process* that describes the relationship between the population size trajectory and the set of local genealogies and (ii) *an observation process* that 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. Population size trajectories are then inferred from sequence data, using these coalescent-based hidden Markov models. In this study, we restrict attention to the state process and present a novel Bayesian approach for inferring population size trajectories from local genealogies. 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 state space of local genealogies for large sample sizes. The pioneering pairwise sequentially Markov coalescent (PSMC) method of Li and Durbin (2011) employed the SMC to infer $N(t)$ from a sample of size 2 ($n = 2$). In this method, time is discretized and the population size trajectory is piecewise constant. Subsequent methods for samples larger than 2 similarly rely on the discretization of time. 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; thus MSMC's 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 piecewise constant trajectory of population sizes.

Gaussian process-based Bayesian inference of population size trajectories has proved 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 Gaussian process (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 piecewise constant or linear framework.

In this article, we overcome the limitations of existing methods—discretizing time, assuming a piecewise constant trajectory, and reporting only point estimates for past population sizes—by introducing a Bayesian nonparametric approach with a 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 approach, 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). To compare our Bayesian GP-based estimation of population size trajectories with a piecewise constant maximum-likelihood-based estimation (*e.g.*, Li and Durbin 2011; Sheehan *et al.* 2013; Schiffels and Durbin 2014), we implement the expectation-maximization (EM) algorithm within our framework and compute the observed Fisher information to obtain confidence intervals of the maximum-likelihood estimates.

Finally, we address a key problem for inference of population size trajectories under sequentially Markov coalescent models: 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).

Methods: SMC′ Calculations

Following notation similar to that in 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 uppercase 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 that records the local genealogies, and let $P = (U, W) = \{(U_i, W_i)\}_{i\in\mathbb{N}}$ represent the chain that 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:

$$\Pr\Big[G_i = g_i, U_i \le u_i, W_i = w_i, S_i \le s \Big| \{g_j, b_j\}_{j=0}^{i-1}, \{u_j, w_j\}_{j=1}^{i-1}\Big]$$

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

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

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

Thus, 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).

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

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

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 piecewise 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 $t_1^i = \infty$.

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:

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

At recombination breakpoint b_i, a new local genealogy g_i is generated that depends on the previous local genealogy g_{i-1} and the population size trajectory $N(t)$ (Figure 1). 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_{new}^i with some lineage, c_i (Figure 2). We then delete the w_i lineage's segment from u_i to t_{del}^i (the coalescent time of lineage w_i). The transition density to a new genealogy at recombination breakpoint b_i is then

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

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

This generative process for local genealogies can result in a *visible transition*, where a genealogy g_i is different from g_{i-1} (Figure 2A), or an *invisible transition*, where g_i is 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 $p_i = (u_i, w_i)$, an invisible transition occurs when $T_{\text{new}}^i \in (u_i, t_{\text{del}}^i)$ 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

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

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

$$\begin{aligned}
&\Pr[G_i = g_{i-1}, p_i = (u_i, w_i) | g_{i-1}, N(t)] \\
&\quad = \frac{1}{l_{i-1}} \Pr[G_i = g_{i-1} | p_i = (u_i, w_i), g_{i-1}, N(t)].
\end{aligned}$$

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	No. 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 no. sequences
l_i	Total tree length of local genealogy g_i
$A^i(t)$	Piecewise 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'
$F^i_{j,k}$	No. free lineages in local genealogy g_i that do not coalesce in the time interval (t^i_{j+1}, t^i_k)
$I^i(t)$	Piecewise constant function that takes values in $\{0, 1, 2\}$ indicating no. ancestral lineages at time t in genealogy g_i where the pruning event would produce a visible transition to g_{i+1}
Discretization	
d	No. change points at which $N(t)$ is estimated
$\mathbf{x} = (x_1, \ldots, x_d)$	Times at which $N(t)$ is estimated

Transition densities averaged over unknown pruning locations

Even though we assume that local genealogies are known, to build inferential frameworks for sequence data in the future, 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.

Visible transitions: To compute the marginal visible transition density to a new genealogy $g_i = \{g_{i-1} \backslash \{t^i_{\text{del}}\} \cup \{t^i_{\text{new}}, (w_i', c_i)\}\}$, 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 $(0, t^i_{\text{del}})$. 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, File S1, and File S2.

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 piecewise 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,

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

For a general piecewise constant function $I^{i-1}(t)$, the marginal visible transition density to a new genealogy is

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

Invisible transitions: To compute the marginal transition probabilities for invisible events, we must 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}. To have an

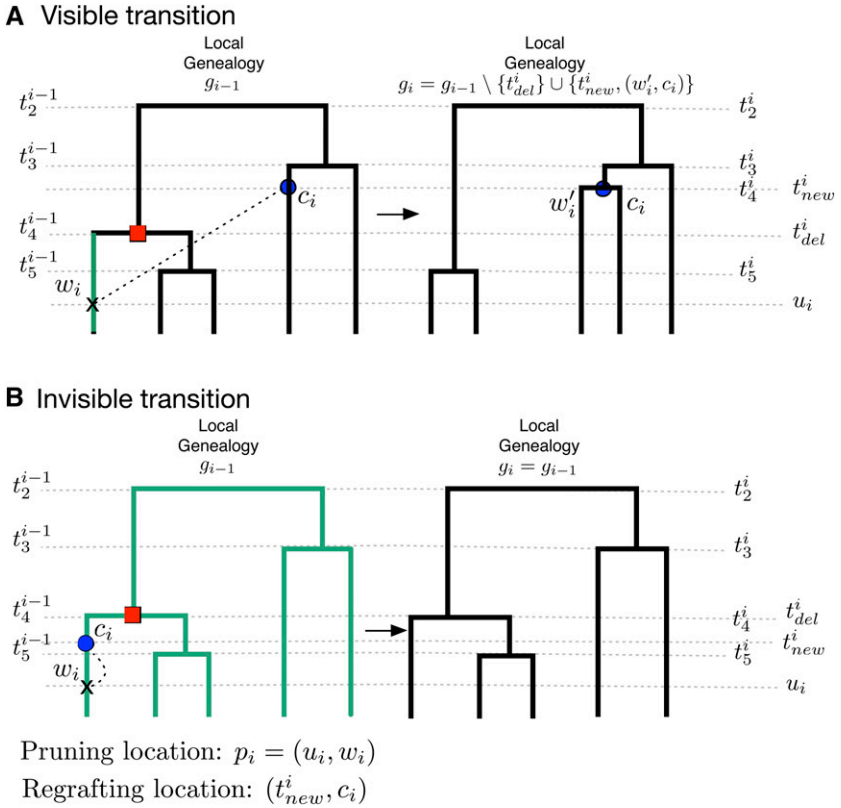


Figure 2 Schematic representation of SMC′ transitions given a recombination breakpoint at location b_i (indicated as an arrow in each panel). (A) Visible transition. We uniformly sample the pruning location p_i from g_{i-1} at time u_i along some branch w_i, and we add a new branch w_i' at u_i and regraft 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 regraft 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.

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 $n \geq j \geq k \geq 2$, 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), with $t^i_{n+1} = 0$. 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 $U_i \in (t^i_{j+1}, t^i_j)$, an invisible transition with new coalescent time $T^i_{\text{new}} \in (t^i_{k+1}, t^i_k)$ 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

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

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 $U_i \in (t^{i-1}_6, t^{i-1}_5)$ and coalescent time T^i_{new} occurs in the interval (t^{i-1}_6, t^{i-1}_5) and this results in an invisible transition has probability

$$\Pr\left[G_i = g_{i-1}, U_i \in (t^{i-1}_6, t^{i-1}_5), T^i_{\text{new}} \in (t^{i-1}_6, t^{i-1}_5) \middle| g_{i-1}, N(t)\right] = \frac{F^{i-1}_{5,5} \int_{t^{i-1}_6}^{t^{i-1}_5} \int_{u_i}^{t^{i-1}_5} (1/N(t)) \exp\left\{ - \int_{u_i}^{t} (A^{i-1}(u)du/N(u)) \right\} dt du_i}{l_{i-1}} \quad (9)$$

$$= \frac{F^{i-1}_{5,5} P^{i-1}_{5,5}}{l_{i-1}}, \quad (10)$$

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

An invisible transition would also result if $U_i \in (t^{i-1}_6, t^{i-1}_5)$ and $T^i_{\text{new}} \in (t^{i-1}_5, t^{i-1}_4)$ along the same (unknown) pruning branch; in Figure 2B, this can happen along three lineages, so $F^{i-1}_{5,4} = 3$ and this event has probability

$$\Pr\left[G_i = g_{i-1}, U_i \in (t^{i-1}_6, t^{i-1}_5), T^i_{\text{new}} \in (t^{i-1}_5, t^{i-1}_4) \middle| g_{i-1}, N(t)\right] = \frac{F^{i-1}_{5,4} \int_{t^{i-1}_6}^{t^{i-1}_5} \exp\left\{ - \int_{u_i}^{t^{i-1}_5} (A^{i-1}(u)du/N(u)) \right\}}{l_{i-1}} \times \int_{t^{i-1}_5}^{t^{i-1}_4} \frac{1}{N(t)} \exp\left\{ - \int_{t^{i-1}_5}^{t} \frac{A^{i-1}(u)du}{N(u)} \right\} dt du_i = \frac{F^{i-1}_{5,4} P^{i-1}_{5,4}}{l_{i-1}}.$$

If we continue considering the cases where $U_i \in (t^{i-1}_6, t^{i-1}_5)$ and $T^i_{\text{new}} \in (t^{i-1}_4, t^{i-1}_3)$ or $T^i_{\text{new}} \in (t^{i-1}_3, t^{i-1}_2)$, we have $F^{i-1}_{5,3} = 2$

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

$$\Pr\Big[G_i = g_{i-1}, U_i \in \big(t_6^i, t_5^i\big) \big| g_{i-1}, N(t)\Big] = \frac{\sum_{k=2}^{5} F_{j,k}^{i-1} P_{j,k}^{i-1}}{l_{i-1}}.$$

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}),

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

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

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

and the probability of coalescing at $T_{\text{new}}^i \in (t_{k+1}^{i-1}, t_k^{i-1})$,

$$1 - q_k^{i-1}.$$

Then,

$$\frac{1}{l_{i-1}} P_{j,k}^{i-1} = \frac{1}{l_{i-1}} Q_j^{i-1} q_{j-1}^{i-1} q_{j-2}^{i-1} \dots q_{k+1}^{i-1} \left(1 - q_k^{i-1}\right)$$

represents the probability that the pruning location is w_i at time $U_i \in (t_{j+1}^{i-1}, t_j^{i-1})$ and the new lineage w_i' coalesces at time $T_{\text{new}}^i \in (t_{k+1}^{i-1}, t_k^{i-1})$ with lineage $c_i = w_i$. Overall, the marginal transition probability to an invisible event is

$$\begin{aligned}
&\Pr[G_i = g_{i-1} | g_{i-1}, N(t)] \\
&\quad = \int_0^{t_2^{i-1}} \Pr\big[G_i = g_{i-1}, u_i | g_{i-1}, N(t)\big] du_i \\
&\quad = \sum_{j=2}^{n} \Pr\Big[G_i = g_{i-1}, U_i \in \Big(t_{j+1}^{i-1}, t_j^{i-1}\Big) \Big| g_{i-1}, N(t)\Big] \\
&\quad = \frac{1}{l_{i-1}} \sum_{j=2}^{n} \sum_{k=2}^{j} F_{j,k}^{i-1} P_{j,k}^{i-1}. \qquad (11)
\end{aligned}$$

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 (including visible and invisible events). Note that our observed data are not sequence data. More specifically, our observed data are

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

Then, the observed data likelihood is

$$\mathcal{L}_{\text{obs}}(\mathbf{Y}; N(t), \rho) = \overbrace{\Pr\Big[g_0 | N(t)\Big] \left[\prod_{i=0}^{m-2} \Pr\Big[g_{i+1} | g_i, N(t)\Big] \right]}^{\text{factors that depend on } N(t)} \times \underbrace{h(L - b_{m-1} | g_{m-1}, \rho) \left[\prod_{i=0}^{m-2} f[s_{i+1} | g_i, \rho] \right]}_{\text{factors that depend on } \rho}, \qquad (13)$$

where $h(L - b_{m-1} | g_{m-1}, \rho)$ 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. Moreover, recombination locations $b_0, b_1, \dots, b_{m-1}$ do not provide information about $N(t)$.

Methods: Inference

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

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

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; further, the fewer observed coalescent times in an interval, the greater the uncertainty is 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 $x_d \le t < \infty$.

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

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

where κ is specified by the user. Schiffels and Durbin (2014) propose discretizing time according to the quantiles of the exponential distribution, $x_j = (-1/\lambda)\log[1 - j/d]$, 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, this equation is not directly applicable here because we use all coalescent events for inference.

In the following sections, we first present our Bayesian nonparametric method and then develop a maximum-likelihood method under a piecewise constant trajectory so we can directly compare an EM-based method to our Bayesian nonparametric method.

Gaussian-process-based Bayesian nonparametric estimation of N(t)

For our Bayesian methodology, we assume the log-Gaussian process prior on the population size trajectory,

$$N(t) = \exp[f(t)],\ f(t) \sim \mathcal{GP}(\mathbf{0}, \mathbf{C}(\tau)), \tag{16}$$

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

$$\begin{aligned}\Pr[N(t), \tau | g_0, \ldots, g_{m-1}] &\propto \Pr[g_0 | N(t)] \\ &\times \left\{ \prod_{i=0}^{m-2} \Pr\left[g_{i+1} | g_i, N(t) \right] \right\} \Pr[N(t)|\tau] \Pr(\tau).\end{aligned} \tag{17}$$

The first two factors on the right side of Equation 17, detailed in Equations 8 and 11, involve integration over $N(t)$, an infinite-dimensional random function (Equation 16). 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,

$$\int_a^b \exp\Big[-f(t)\Big]dt \approx \sum_{j=i}^{k} \exp\Big[-f_j^*\Big]\Delta_j, \tag{18}$$

for $x_i < a < x_{i+1} < \ldots < x_{k-1} < b < x_k$, $\Delta_i = x_{i+1} - a$, $\Delta_k = b - x_{k-1}$, and $\Delta_j = x_{j+1} - x_j$ for $i < j < k-1$. 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 $x_1 = 0 < x_2 < \ldots < x_d$, with $x_d = \max(t_1^0, \ldots, t_1^{m-1})$, the maximum time to the most common ancestor observed in the sequence of local genealogies, and approximate $N(t)$ by a piecewise linear function evaluated at $(x_1^*, x_2^*, \ldots, x_d^*)$.

We condition on the set of m local genealogies $g_0, \ldots, g_{m-1}$ (assuming pruning locations are not known) to generate posterior samples for the vector $\mathbf{f}^* = [\log N(x_1^*), \ldots, \log N(x_d^*)]$ and τ and use these posterior samples to infer $N(t)$ at

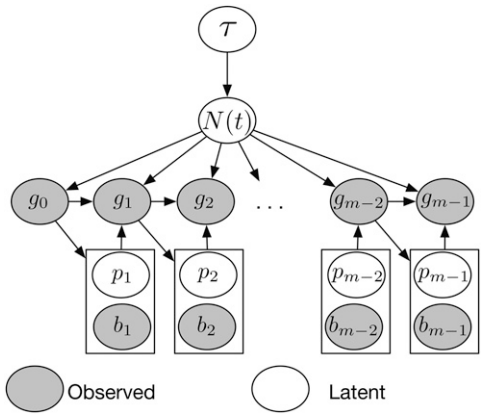


Figure 3 Structure of our Bayesian model for inferring population size trajectories from a realization of the SMC′ process at recombination breakpoints. Hyperparameter τ controls the smoothness of the log-Gaussian process prior on $N(t)$. Local genealogies depend on $N(t)$ and form a Markov chain of degree 1. 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}.

$t \in (x_1^*, \ldots, x_d^*)$, where $x_i^* = (x_i + x_{i+1})/2$. Updating the vector $\mathbf{f}$ and τ separately is not recommended because of their strong dependency (Lan *et al.* 2015). Therefore, we update $(\mathbf{f}, \tau)$ jointly in an MCMC sampling algorithm, using 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. The splitHMC method relies on Hamiltonian dynamics to propose a new state of the model parameters jointly with a higher acceptance rate than simple methods such as random-walk Metropolis (Neal 2009). splitHMC 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-likelihoods of the observed data are approximated via sums of the form in Equation 18. We approximate the score function $\nabla\mathcal{L}_{\text{obs}}(\mathbf{Y}; \mathbf{f}^*)$ with respect to $\mathbf{f}^*$ by applying Fisher's identity,

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

where, at each iteration in the MCMC, expectation is calculated using the current value of $\mathbf{f}^*$ (see *Appendix*).

Alternatively, one can update $N(t)$ in the MCMC algorithm, using the 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 the elliptical slice sampler over the split Hamiltonian Monte Carlo is purely computational (the elliptical slice sampler does not require calculation of the score function).

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 8 and 11 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 implement an 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 $\{t^i_{\text{new}}\}_{i\in\mathcal{I}}$, where $\mathcal{I} \subset \{1, 2, \ldots, m-1\}$ denotes the set of indexes of invisible transitions.

The complete data log-likelihood is then

$$\mathcal{L}_c(\mathbf{Y}_c; \mathbf{N}) := \log \Pr[g_0|N(t)] + \sum_{i=1}^{m-1} \log Pr\big[p_i = (u_i, w_i), t^i_{\text{new}}, c_i | g_{i-1}, N(t)\big]. \tag{19}$$

The EM algorithm starts by initializing the population size trajectory to a piecewise 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

$$\mathbf{N}^k = \arg\max_{\mathbf{N}} E_{\mathbf{N}^{k-1}}[\mathcal{L}_c(\mathbf{Y}_c; \mathbf{N})|\mathbf{Y}]. \tag{20}$$

The conditional expectation in Equation 20 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 $x^i_{j-1} < t^{i+1}_{\text{new}} \le x^i_j$. For ease of notation, we denote the number of time intervals $|\mathbf{x}^i|$ by $D = d + n - 2$. Let

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

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

$$\log \Pr[g_0|N(t)] = -\sum_{j=1}^{D}\Big\{a^0_j \log N\big(x^0_{j+1}\big)\Big\} - \sum_{j=1}^{D}\left\{\frac{A^0\big(x^0_{j+1}\big)\Big[A^0\big(x^0_{j+1}\big) - 1\Big]\big(x^0_{j+1} - x^0_j\big)\exp\Big[-\log N\big(x^0_{j+1}\big)\Big]}{2}\right\}. \tag{21}$$

Let

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

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

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

Then, the augmented transition density can be expressed as

$$\begin{aligned} &\log \Pr\big[p_i = (u_i, w_i), t^i_{\text{new}}, c_i | g_{i-1}, N(t)\big] \\ &= \log \Pr\Big[p_i = (u_i, w_i), t^i_{\text{new}}, c_i, \mathbf{z}^i, \boldsymbol{\Delta}^i | g_{i-1}, N(t)\Big] \\ &= -\log l_{i-1} - \sum_{j=1}^{D}\big\{z^{i-1}_j \log N\big(x^{i-1}_{j+1}\big)\big\} \\ &\quad - \sum_{j=1}^{D}\Big\{A^{i-1}\big(x^{i-1}_{j+1}\big)\Delta^{i-1}_j \exp\big[-\log N\big(x^{i-1}_{j+1}\big)\big]\Big\}, \end{aligned} \tag{22}$$

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 $E[\mathbf{z}^i_j|\mathbf{Y}]$ and $E[\boldsymbol{\Delta}^i_j|\mathbf{Y}]$. The details of these calculations are in the *Appendix*.

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

$$\begin{aligned} \hat{\mathbf{I}}_Y\big[\hat{\mathbf{N}}\big] &= \mathrm{E}_{\hat{\mathbf{N}}}\big[-\hat{\mathbf{H}}\mathcal{L}_c\big(\mathbf{Y}_c; \hat{\mathbf{N}}\big)\big|\mathbf{Y}\big] \\ &\quad - \mathrm{E}_{\hat{\mathbf{N}}}\Big[\nabla\mathcal{L}_c\big(\mathbf{Y}_c; \hat{\mathbf{N}}\big)\nabla\mathcal{L}_c\big(\mathbf{Y}_c; \hat{\mathbf{N}}\big)'\Big|\mathbf{Y}\Big], \end{aligned}$$

where $\nabla\mathcal{L}_c(\mathbf{Y}_c; \hat{\mathbf{N}})$ is the gradient and $\mathbf{H}\mathcal{L}_c(\mathbf{Y}_c; \hat{\mathbf{N}})$ 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 File S7.

Data availability

The R code for all simulation studies and analysis of sequence data conducted in this article are publicly available at http://ramachandran-data.brown.edu/.

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 File S3 for details of these simulations. We assumed that all individuals were sampled at time $t = 0$.

Table 2 Simulated demographic scenarios

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	$N(t) = \begin{cases} 1, & \text{for } t \in (0, 0.3), \\ 0.1, & \text{for } t \in (0.3, 0.5), \\ 1, & \text{for } t \in (0.5, \infty). \end{cases}$

The argument t denotes time measured in units of N_0 generations.

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

$$\text{SRE} = \sum_{i=1}^{K} \frac{\left|\hat{N}(x_i) - N(x_i)\right|}{N(x_i)}, \tag{23}$$

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

$$\text{MRW} = \sum_{i=1}^{K} \frac{\left|\hat{N}_{\text{up}}(x_i) - \hat{N}_{\text{low}}(x_i)\right|}{KN(x_i)}, \tag{24}$$

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

$$\text{ENV} = \frac{\sum_{i=1}^{K} I\left(\hat{N}_{\text{up}}(x_i) \le N(x_i) \le \hat{N}_{\text{low}}(x_i)\right)}{K}. \tag{25}$$

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 consider only 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 in Figure 5, Figure 6, and Figure 7. 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).

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 4, B and 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 3) 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 number of change points from 5 to 10 for $\kappa = 10$ (Figure 4, B and C). EM grid sensitivity is persistent even when the number of genealogies increases; Figure S2 in File S4 shows that the best definition of change points when our data consist of 1000 local genealogies of 100 individuals is 10 evenly distributed change points.

Comparing methods for estimating N(t)

Figure 5 shows the estimated population size trajectories when the number of samples is two 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 piecewise 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 piecewise constant trajectory of 5 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

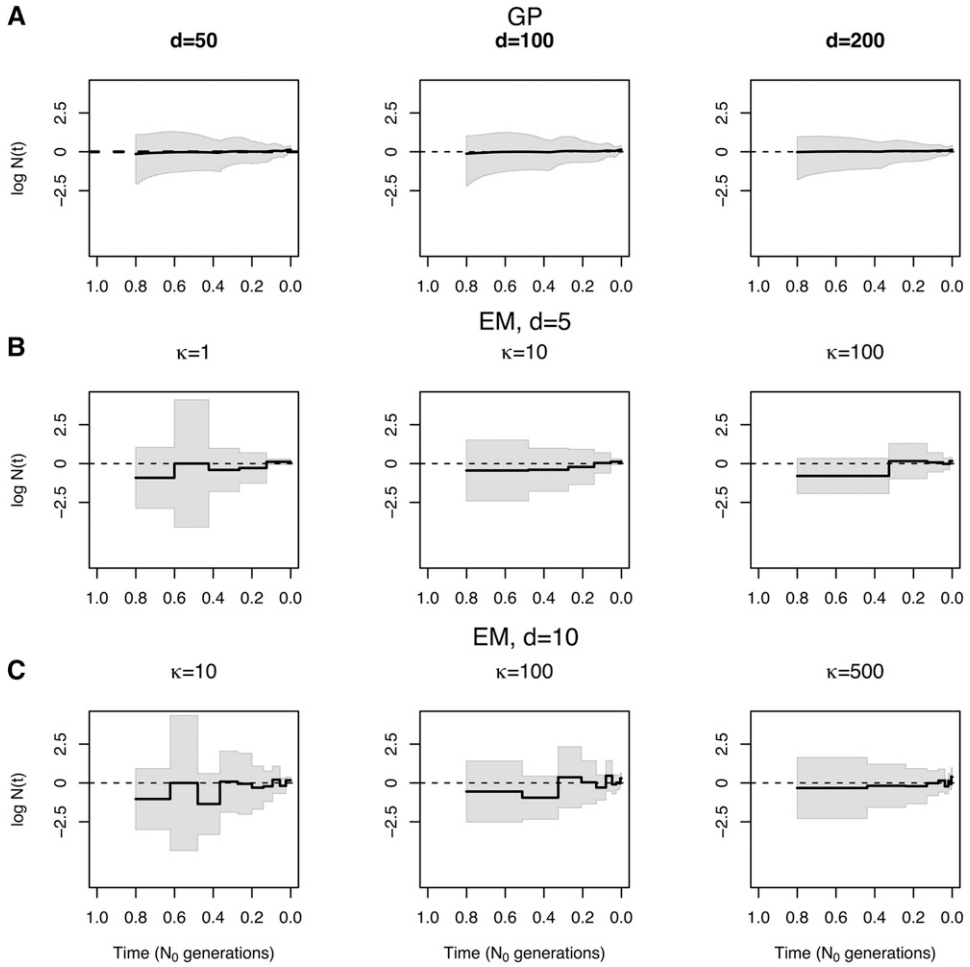


Figure 4 Sensitivity to parameter discretization. Population size trajectories estimated from one simulated genealogy ($m = 1$) of 100 individuals with a constant population size are compared. We show true trajectories as dashed lines. (A) Bayesian GP estimates at $d =$ 50, 100, and 200 equally spaced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

each plot in Figure 5, which indicate the uncertainty expected in 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 size model in Figure 5A, our Bayesian GP considerably outperforms our EM estimates. This is not surprising since *a priori* $\log N(t)$ has mean 0 in our Bayesian approach (Equation 16). Moreover, EM confidence intervals cover the truth only ~30% of the time, while the GP method covers 100% of the truth (Table 4A). Despite placing a mean-0 prior on $\log N(t)$, the Bayesian GP method accurately recovers sudden changes as shown in the bottleneck scenario. Although our Bayesian GP prior on $\log N(t)$ is Brownian motion (which is not differentiable at any point), 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 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.

Figure 6 and Figure 7 show our estimates when $n = 20$ and $n = 100$ (Table 4, B and C, gives performance statistics). 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.

Sampling more individuals vs. sequencing more loci

Figure 5, Figure 6, and Figure 7 show our estimates for $n = 2$, 20, and 100 sampled individuals across varying numbers of loci. Performance of EM estimates depends strongly on the definition of the grid, so we focus here on the Bayesian GP estimates. We find that increasing the number of loci decreases uncertainty of our estimates and allows us to infer $N(t)$ farther back in time. Increasing the number of samples

Table 3 Summary statistics for simulation results depicted in Figure 4

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**

SRE is the sum of relative errors (Equation 23), MRW is the mean relative width of the 95% BCI (Equation 24), and ENV is the envelope measure (Equation 25). Values in boldface type indicate best performance.

does not necessarily increase the performance of our GP estimates (File S6). For example, under the bottleneck scenario, we are able to detect the bottleneck fairly accurately even for two samples with $m = 1000$ local genealogies. This is because most TMRCAs observed under the bottleneck scenario occur during the bottleneck (Figure 5, Figure 6, and Figure 7), regardless of the sample size. In contrast, in our exponential growth scenario, increasing the number of samples from $n = 2$ to $n = 100$ improves accuracy: point estimates are closer to the truth (SRE in Table 4, A–C) and credible intervals cover the truth completely (ENV of 100%).

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

Under the SMC′, 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 sequences 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 carries information about $N(t)$. For a single locus, the set of coalescent times provides sufficient statistics for inferring $N(t)$ (see *Proof* 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* (tree with no labels but ranked coalescent events). For a single locus, the set of coalescent times together with the ranked tree shape corresponds to a realization of Tajima's

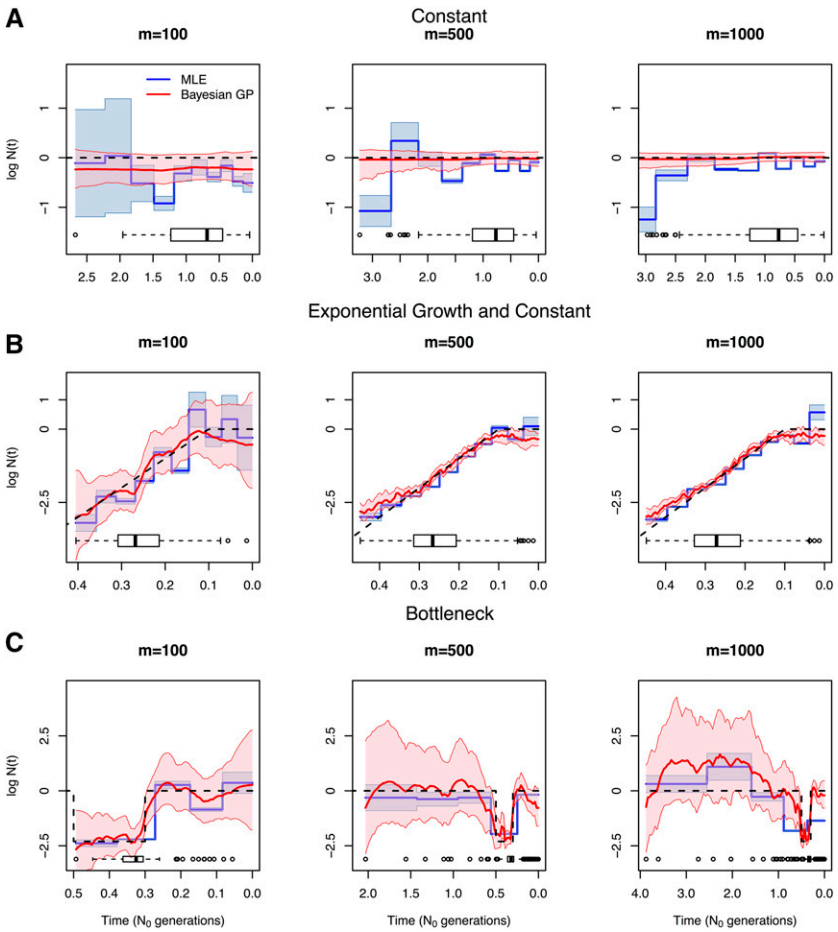


Figure 5 Inference of population size trajectories $N(t)$ for a pair of individuals ($n = 2$). Simulated data under constant population size (A), exponential and constant trajectory (B), and a bottleneck (C). We show estimates from $m = 100$, $m = 500$, and $m = 1000$ local genealogies. Dashed lines show the true trajectories, blue lines and light blue areas represent EM point estimates and 95% confidence areas, and red lines and pink 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 Figure 5

A. Simulations with $n = 2$

Simulation and method	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. Simulations 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. Simulations 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. GP	**90.27**	**44.14**	**42.68**	6.98	2.62	1.74	**100.0**	**94.7**	**96.0**

SRE is the sum of relative errors calculated as in (23), MRW is the mean relative width of the 95% BCI as Const: Constant population simulation scenario. Exp: Exponential growth followed by constant population size simulation scenario and Bottle: Population bottleneck. defined in (24), and ENV is the envelope measure calculated as in (25). Values in boldface type indicate best performance for each demographic model and sample size.

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 10 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

$$P(H_n) = \frac{2^{n-c-1}}{(n-1)!}, \tag{26}$$

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

We defined transition densities in terms of coalescent times and $F_{i,j}$ quantities (see *Methods: SMC′ calculations*). The set of all $F_{i,j}$ quantities from a local genealogy forms a triangular matrix: an F matrix. We show that (i) F matrices are in bijection with ranked tree shapes and (ii) the set of local Tajima's genealogies has sufficient statistics for inferring $N(t)$ under the SMC′ model (see *Appendix*). These observations are crucial for inferring $N(t)$ from sequence data directly. Coalescent-based inference from sequence data relies 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 7936 possible ranked tree shapes.

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 *ARG*weaver (Rasmussen *et al.* 2014) to obtain a sample path of local genealogies for the two populations (YRI and CEU). The

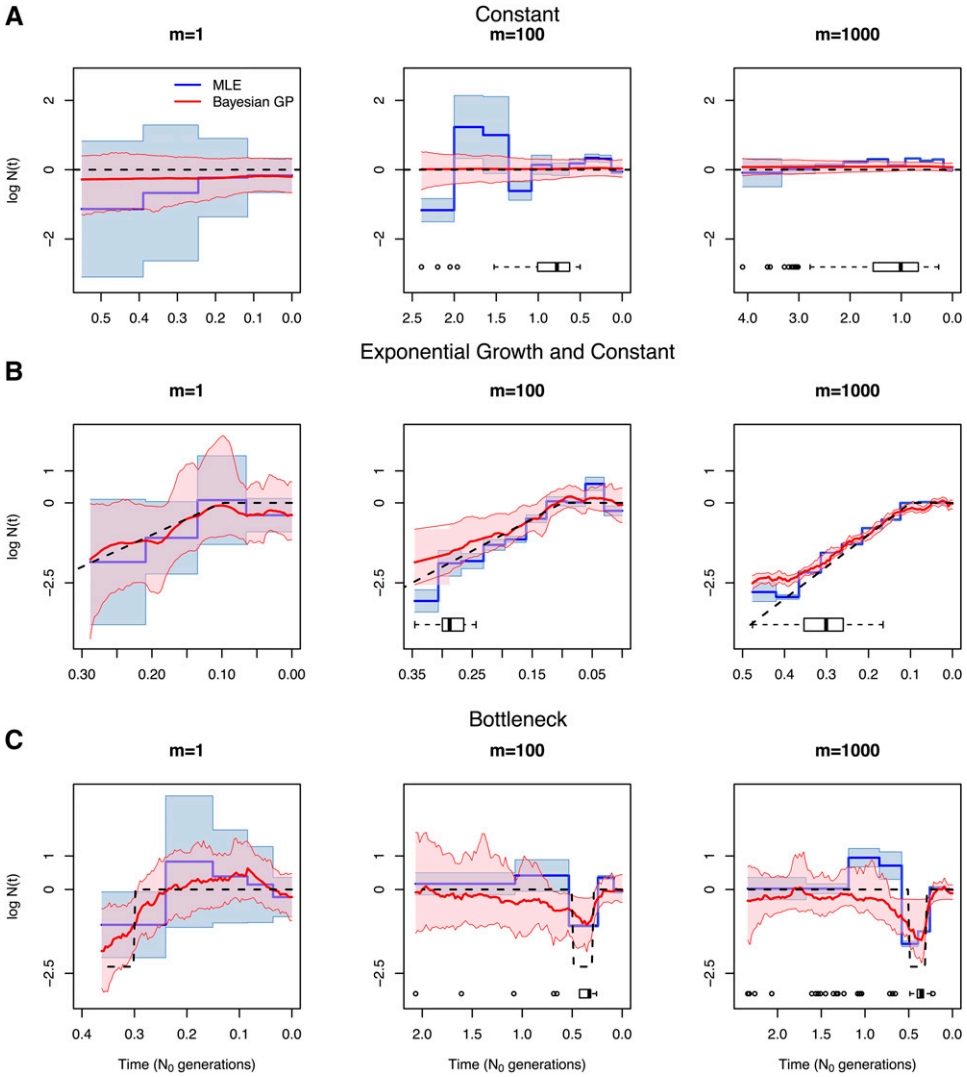


Figure 6 Inference of population size trajectories $N(t)$ for $n = 20$. Simulated data under constant population size (A), exponential and constant trajectory (B), and a bottleneck (C). We show estimates from $m = 1$ genealogy, $m = 100$ local genealogies, and $m = 1000$ local genealogies. Dashed lines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

parameters used were 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) (File S5). We note that *ARG*weaver 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 *ARG*weaver set of local genealogies is 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 ∼100 KYA, 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 bottleneck that starts ∼100 KYA and ends ∼30 KYA in the European population. These results are consistent with previously published results (Gronau *et al.* 2011; Li and Durbin 2011; Rasmussen *et al.* 2011; Sheehan *et al.* 2013; Schiffels and Durbin 2014). In File S5, Figure S4A, we show the estimates of $\log N(t)$ instead of $N(t)$ and time measured in units of N_0 generations (as in Figure 5, Figure 6, and Figure 7). We note that this two-step procedure of inferring local genealogies with *ARG*weaver and then using our method introduces biases and ignores genealogical uncertainty. In File S5, we correct for some of the bias caused by using this two-step procedure and show that our inferred population size trajectory remains valid for the recent past.

Assessing the effect of using genealogies inferred with ARGweaver

We simulated sequence data, used *ARG*weaver for inferring a set of local genealogies, and used our method on those genealogies to obtain estimates of $\log N(t)$. To this end, we took the sequences of the first 1000 local genealogies of $n = 20$ individuals simulated with MaCS as described in section 3 of File S5. We then generated the sequence lengths (s_i, for the locus corresponding to g_{i-1}) as in Equation 5,

$$s_i \sim \text{Exponential}(\rho \times l_{i-1} \times N_0),$$

where l_{i-1} is the tree length of g_{i-1} in units of N_0 generations and N_0 is the current population size. In our simulations, we set $N_0 = 20{,}000$, $\rho = 1.8 \times 10^{-8}$. To simulate sequence data

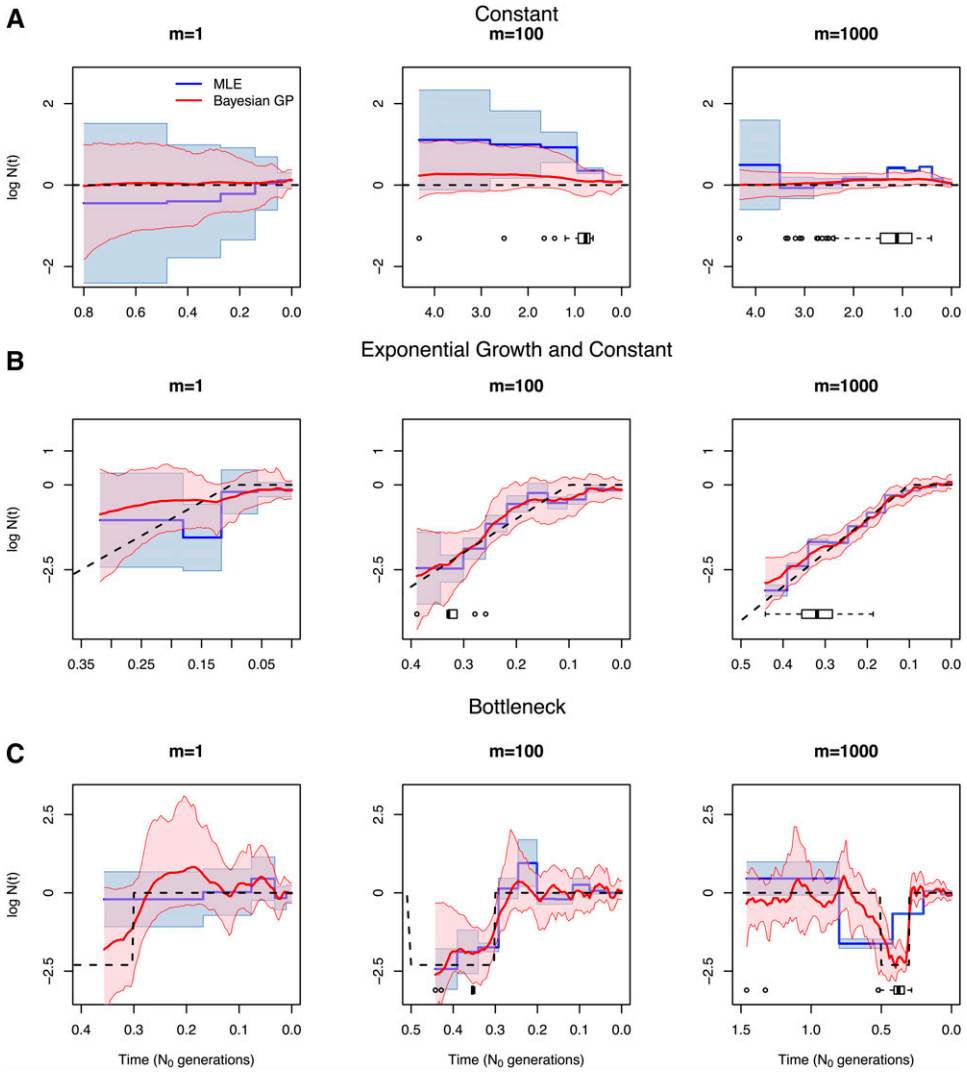


Figure 7 Inference of population size trajectories $N(t)$ for $n = 100$. Simulated data under constant population size (A), exponential and constant trajectory (B), and bottleneck (C). We show estimates from $m = 1$ genealogy, $m = 100$ local genealogies, and $m = 1000$ local genealogies. Dashed lines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

of length s_i over genealogy g_{i-1}, we used Seq-Gen (Rambaut and Grassly 1997) implemented in the R package phyclust (Chen 2011) from the Jukes–Cantor mutation model (Jukes and Cantor 1969) with mutation rate $\mu = 2 \times 1.8 \times 10^{-8}$. We then used *ARG*weaver to infer a sample of local genealogies with the same corresponding parameters and with 200 change points for discretization of time. Figure 9 shows three estimations of effective population size trajectories for our three simulation scenarios. Figure 9, A–C, left, shows our GP-based estimates from 1000 simulated genealogies from MaCS; Figure 9, A–C, center, shows our GP-based estimates from a realization of local genealogies obtained from *ARG*weaver; and Figure 9, A–C, right, shows our GP-based estimates correcting the number of lineages used in our calculations, replacing $A^i(t)$ by $A^i(t) - 1$ in our likelihood calculations. We find that our estimates are not only noisier using this approach but also biased.

Discussion

In this article, 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 (Figure 5, Figure 6, and Figure 7). We apply our method to genealogies estimated using *ARG*weaver (Rasmussen *et al.* 2014) for European and African samples in the 1000 Genomes; this application to real data recovers the known features of the out-of-Africa bottleneck (Figure 8).

Several recent approaches have emerged for inferring population size trajectories from multiple whole-genome sequences using the 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 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 piecewise constant or a piecewise linear function—the Bayesian

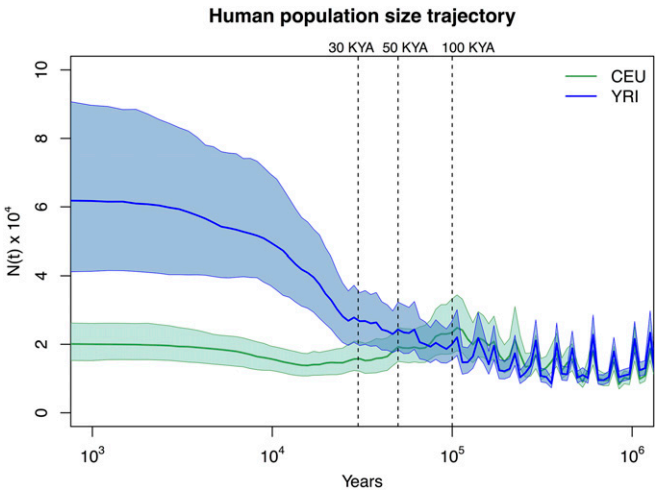


Figure 8 Inference of human population size trajectories $N(t)$ for $n = 10$. Green solid line and green areas represent the posterior median and 95% BCI for the European population (CEU) and blue solid line and blue areas represent the posterior median and 95% BCI for the Yoruban population (YRI). Time 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.

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 (Figure 5, Figure 6, and Figure 7). For example, in simulation, our EM approach incorrectly detects the initial period of the simulated bottleneck ($\sim 0.8N_0$ instead of $0.5N_0$ generations ago) with narrow confidence intervals (Figure 7C).

Using Bayesian GP for inferring population size trajectories offers many advantages over the EM approach. 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 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 the 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 (Figure 5, Figure 6, and Figure 7). However, resolution into the past from contemporaneous sequences highly depends on the actual population size trajectory $N(t)$.

We used *ARG*weaver (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 European effective population size trajectories (Figure 8). We find an out-of-Africa bottleneck that began ~100 KYA and ended ~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 *ARG*weaver at 200 time intervals, so our GP-based approach cannot fully detect sudden changes that may occur between the discretized times. In addition, *ARG*weaver assumes an SMC prior model on local genealogies and our GP-based method assumes the SMC′ process; the lack of invisible recombination events in *ARG*weaver's genealogies will bias inference (as shown in simulated data in Figure 9).

The natural next extension for our method presented in this study is to infer $N(t)$ from sequence data directly and not from

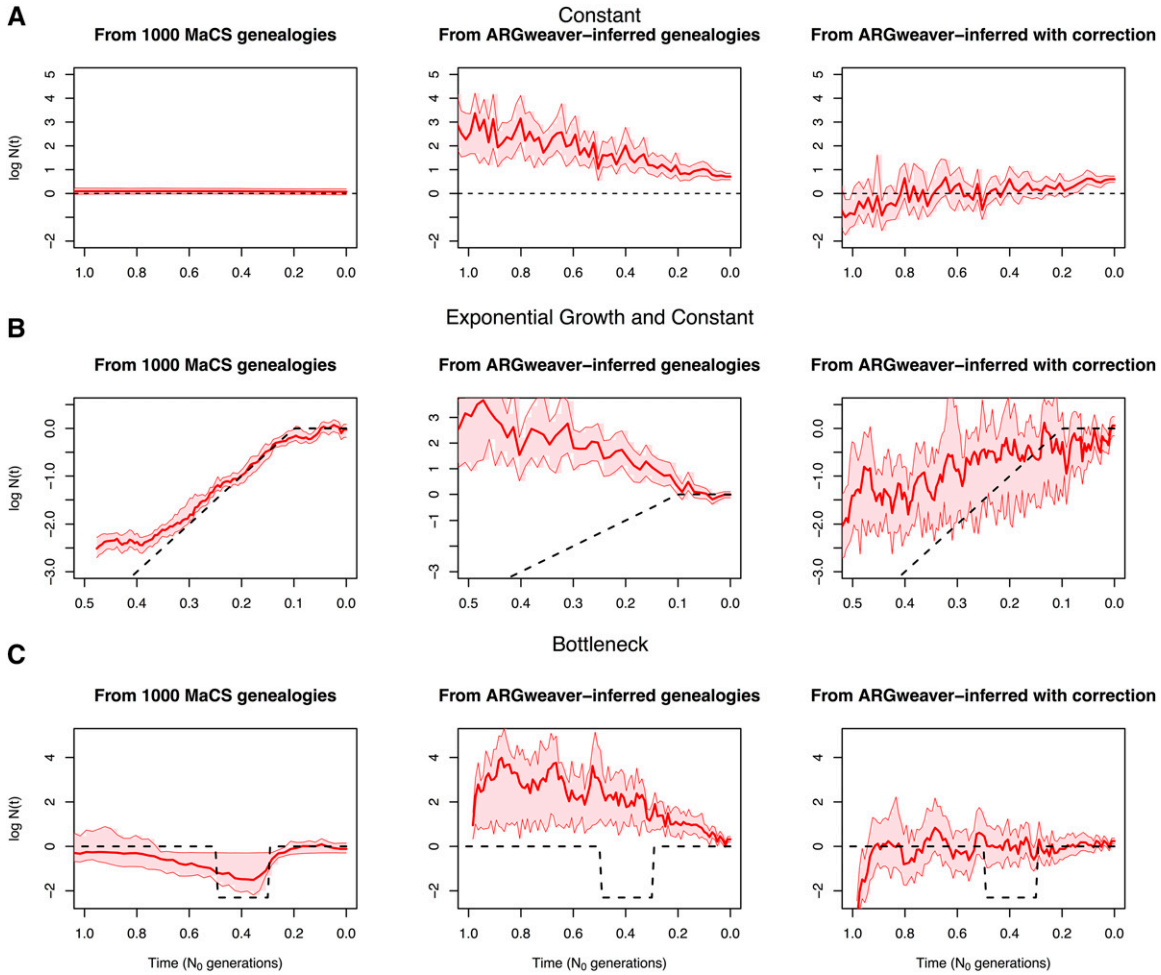


Figure 9 Assessing the effect of a two-step inferential approach. (A–C) Left, our GP-based estimates when 1000 local genealogies are known; center, our GP-based estimates when *ARG*weaver estimated local genealogies are used for inference; right, our corrected GP-based estimates when *ARG*weaver estimated local genealogies are used for inference.

the set of local genealogies. Our MCMC approach allows us to 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 *ARG*weaver 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 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 trade-offs 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.

Acknowledgments

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. We also thank two referees for suggestions that improved this article. J.A.P. acknowledges scholarship from Consejo Nacional de Ciencia y Tecnología Mexico to pursue her research work. This research is supported in part by National Science Foundation 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

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

Chen, W., 2011 Overlapping codon model, phylogenetic clustering, and alternative partial expectation conditional maximization algorithm. Ph.D. Thesis, Iowa State University, Ames, IA.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39: 1–38.

Disanto, F., and T. Wiehe, 2013 Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. Math. Biosci. 242: 195–200.

Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph, pp. 257–270 in *Progress in Population Genetics and Human Evolution* (IMA Volumes in Mathematics and Its Applications), Vol. 87, edited by P. Donnelly and S. Tavaré. Springer-Verlag, New York.

Gronau, I., M. J. Hubisz, B. Gulko, C. G. Danko, and A. Siepel, 2011 Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43: 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. Oxf. Surv. Evol. Biol. 7: 1–44.

Jukes, T. H., and C. R. Cantor, 1969 *Evolution of Protein Molecules*, pp. 21–132. Academic Press, New York.

Kingman, J., 1982 The coalescent. Stoch. Proc. Appl. 13: 235–248.

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

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

Louis, T. A., 1982 Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. B 44: 226–233.

Marjoram, P., and J. Wall, 2006 Fast "coalescent" simulation. BMC Genet. 7: 16.

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

Minin, V. N., E. W. Bloomquist, and M. A. Suchard, 2008 Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25: 1459–1471.

Murray, I., R. P. Adams, and D. J. C. MacKay, 2010 Elliptical slice sampling, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (AISTATS), Chia Laguna Resort, Sardinia, Italy, Vol. 9, pp. 541–548.

Neal, R. M., 2009 MCMC using Hamiltonian dynamics, pp. 113–162, edited by S. Brooks, A. Gelman, G. L. Jones, and X. -L. Meng. Hall, London/New York.

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

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

Rambaut, A., and N. C. Grassly, 1997 Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13: 235–238.

Rasmussen, M., X. Guo, Y. Wang, K. E. Lohmueller, S. Rasmussen *et al.*, 2011 An aboriginal Australian genome reveals separate human dispersals into Asia. Science 334: 94–98.

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

Sainudiin, R., T. Stadler, and A. Véber, 2014 Finding the best resolution for the Kingman-Tajima coalescent: theory and applications. J. Math. Biol. 70: 1207–1247.

Schiffels, S., and R. Durbin, 2014 Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46: 919–925.

Shahbaba, B., S. Lan, W. Johnson, and R. Neal, 2014 Split Hamiltonian Monte Carlo. Stat. Comput. 24: 339–349.

Sheehan, S., K. Harris, and Y. S. Song, 2013 Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics 194: 647–662.

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

Wilton, P. R., S. Carmi, and A. Hobolth, 2015 The SMC′ is a highly accurate approximation to the ancestral recombination graph. Genetics 200: 343–355.

Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. 55: 248–259.

Communicating editor: Y. S. Song

Appendix

Discretization

For both our Bayesian method and our EM method, we assume that $N(t)$ is a piecewise linear (or piecewise constant) function with d change points. Let $\mathbf{x}^i = \{x_1^i, x_2^i, \ldots, x_{d+n-1}^i\}$ 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 19).

Let $\tilde{F}_{k,j}^i$ 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_k^i, x_{j+1}^i). Note that the indexes here are in increasing order, $k \le j$. Similarly, let $(1/l_i)\tilde{P}_{k,j}^i$ denote the probability that U_i (the pruning time along genealogy i) occurs in (x_k^i, x_{k+1}^i) and the self-coalescing event occurs at time t_{new}^i in (x_j^i, x_{j+1}^i). That is,

$$\tilde{P}_{k,j}^i = \begin{cases} \dfrac{1}{A^i\left(x_{j+1}^i\right)}\left(\Delta_j^i - \tilde{Q}_j^i\right) & k = j \\ \tilde{Q}_k^i \tilde{q}_{k+1}^i \tilde{q}_{k+2}^i \cdots \tilde{q}_{j+1}^i \left(1 - \tilde{q}_j^i\right) & k < j, \end{cases} \tag{A1}$$

where

$$\frac{1}{l_i}\tilde{Q}_k^i = \frac{1}{l_i}\frac{N\left(x_{k+1}^i\right)}{A^i\left(x_{k+1}^i\right)}\left[1 - \tilde{q}_k^i\right] \tag{A2}$$

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

$$\tilde{q}_k^i = \exp\left\{-\frac{A^i\left(x_{k+1}^i\right)\Delta_k^i}{N\left(x_{k+1}^i\right)}\right\}.$$

Expectation-Maximization Algorithm

E step

Equations 21 and 22 show that for the E-step, the only expectations we need are $E[z_j^i|\mathbf{Y}]$ and $E[\Delta_j^i|\mathbf{Y}]$ We compute these expressions as follows:

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

For $i \in \mathcal{I}$

$$\begin{aligned} E\left[\Delta_j^i\middle|Y\right] = {} & \left(x_{j+1}^i - x_j^i\right)\frac{\sum_{k=1}^{j-1}\sum_{l=j+1}^{D}\tilde{F}_{k,l}^i\tilde{P}_{k,l}^i}{\sum_{j=1}^{D}\sum_{k=j}^{D}\tilde{F}_{k,j}^i\tilde{P}_{k,j}^i} \\ & + \int_{x_j^i}^{x_{j+1}^i}\left(x_{j+1}^i - u\right)\exp\left\{-\frac{\left(x_{j+1}^i - u\right)A^i\left(x_{j+1}^i\right)}{N\left(x_{j+1}^i\right)}\right\}du\,\frac{\sum_{k=j+1}^{D}\tilde{F}_{j,k}^i\left(\tilde{P}_{j,k}^i/\tilde{Q}_j^i\right)}{\sum_{j=1}^{D}\sum_{k=j}^{D}\tilde{F}_{k,j}^i\tilde{P}_{k,j}^i} \\ & + \int_{x_j^i}^{x_{j+1}^i}\int_{u}^{x_{j+1}^i}(t-u)\frac{1}{N\left(x_{j+1}^i\right)}\exp\left\{-\frac{(t-u)A^i\left(x_{j+1}^i\right)}{N\left(x_{j+1}^i\right)}\right\}dtdu\,\frac{\tilde{F}_{j,j}^i}{\sum_{j=1}^{D}\sum_{k=j}^{D}\tilde{F}_{k,j}^i\tilde{P}_{k,j}^i} \\ & + \int_{x_j^i}^{x_{j+1}^i}\left(t - x_j^i\right)\exp\left\{-\frac{\left(t - x_j^i\right)A^i\left(x_{j+1}^i\right)}{N\left(x_{j+1}^i\right)}\right\}dt\,\frac{\sum_{k=1}^{j-1}\tilde{F}_{k,j}^i\left(\tilde{P}_{k,j}^i/\left(1-\hat{q}_j^i\right)\right)}{\sum_{j=1}^{D}\sum_{k=j}^{D}\tilde{F}_{k,j}^i\tilde{P}_{k,j}^i}, \end{aligned}$$

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

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

Then,

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

where

$$\delta_j^i = (x_{j+1}^i - x_j^i)\left[\frac{\sum_{k=1}^{j-1} I^i\left(x_{k+1}^i\right)\tilde{Q}_k^i \prod_{l=k+1}^{D_i-1}\left[\hat{q}_l^i\right]^{y_l^i}}{\sum_{k=1}^{D} I^i\left(x_{k+1}^i\right)\tilde{Q}_k^i \prod_{l=k+1}^{D}\left[\hat{q}_l^i\right]^{y_l^i}}\right] \tag{A3}$$

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

M step

Now, for the kth iteration of the algorithm and maximizing the complete data log-likelihood (Equation 19), we have

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

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_j^i, x_{j+1}^i).

Observed Score Function for Split Hamiltonian Monte Carlo

Our Bayesian approach relies on splitHMC sampling 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 $\nabla\mathcal{L}_{\text{obs}}$ is

$$\begin{aligned}(\nabla\mathcal{L}_{\text{obs}})_l = & -\sum_{j=1}^{D-1} a_j^0 1_{l,j}^0 + \frac{1}{2}\sum_{j=1}^{D-1} A^0(x_{j+1}^0)\left[A^0(x_{j+1}^0) - 1\right](x_{j+1}^0 - x_j^0)1_{l,j}^0 \exp[-\log N_l] \\ & -\sum_{i=0}^{m-2}\sum_{j=1}^{D-1} E\left[z_j^i|\mathbf{Y}\right]1_{l,j}^0 + \sum_{i=0}^{m-2}\sum_{j=1}^{D-1} A^i(x_{j+1}^i)E\left[\Delta_j^i|\mathbf{Y}\right]1_{l,j}^i \exp[-\log N_l].\end{aligned} \tag{A4}$$

Sufficient Statistics Under SMC′

Here, we first formally show that under the standard coalescent and for a single locus, the set of coalescent times has sufficient statistics for inferring $N(t)$; that is, information about the topology is irrelevant for inference of $N(t)$. We then investigate the properties of our F quantities (see *Methods: SMC′ Calculations*) needed for the calculation of the transition densities

(Equation 11) and show through a series of propositions that the F quantities and coalescent times are the sufficient statistics for inferring $N(t)$ under the SMC′ process.

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 *Methods: SMC′ Calculations* and with the following properties:

1. $F_{i,1} = 0$ for all $i = 1, \ldots, n$ (The first column contains 0's 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 lose 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

$$F_{n,j-1} = \begin{cases} F_{n,j} - 2, & \text{with probability } p = \dfrac{\binom{F_{n,j}}{2}}{\binom{j}{2}}, \\ F_{n,j} - 1, & \text{with probability } p = \dfrac{F_{n,j}\left(j - F_{n,j}\right)}{\binom{j}{2}}, \\ F_{n,j}, & \text{with probability } p = \dfrac{\binom{j - F_{n,j}}{2}}{\binom{j}{2}}. \end{cases}$$

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

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

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, and 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 $(0,\ 0,\ 2,\ 3,\ 5)$.

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,

$$\begin{aligned} \Pr(F) &= \Pr(F_{\cdot,n}) \prod_{j=1}^{n-1} \Pr\left(F_{\cdot,n-j} \middle| F_{\cdot,n-j+1}\right) \\ &= \prod_{j=2}^{n-2} \Pr\left(F_{\cdot,n-j} \middle| F_{\cdot,n-j+1}\right), \end{aligned}$$

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

$$\Pr\left(F_{\cdot,n-j} \middle| F_{\cdot,n-j+1}\right) = \Pr\left(d_{n-j} | F_{\cdot,n-j+1}\right) \Pr\left(F_{n-j:n-1,n-j} \middle| d_{n-j}, F_{\cdot,n-j+1}\right). \tag{A5}$$

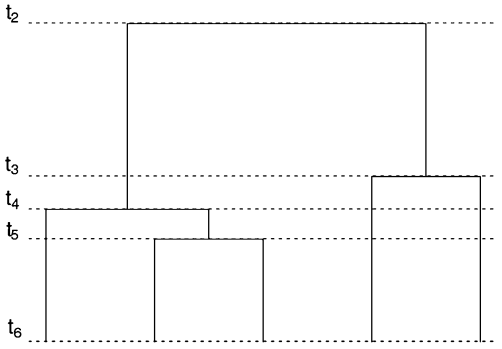
That is, the conditional probability of the $(n-j)$th column of F given the $(n-j+1)$th column of F is the product of the conditional probability of the last element of the $(n-j)$th column and the conditional probability of the rest of the $(n-j)$th column. When $d_{n-j} = 2$, the rest of the column 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 $1/|\nu_{n-j+1}|$ (property 9 of the F matrix) and when $d_{n-j} = 2$, the rest of the $n-j$th column has probability $1 \Big/ \binom{|\nu_{n-j+1}|}{2}$ (property 10 of the F matrix). Then rewriting Equation A5, we have

$$\Pr\left(F_{\cdot,n-j} \middle| F_{\cdot,n-j+1}\right) = \Pr\left(d_{n-j+1} | F_{\cdot,n-j+1}\right) \left(\frac{1}{|\nu_{n-j+1}|}\right)^{1_{\{d_{n-j}=1\}}} \left(\frac{1}{\binom{|\nu_{n-j+1}|}{2}}\right)^{1_{\{d_{n-j}=0\}}}; \tag{A6}$$

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

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

and



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

Figure A1 Ranked tree shape for $n = 5$.

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

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

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

If we continue expanding the expressions, we get

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

■

Note that the entries of the F matrix correspond to the same quantities needed to express the transition density of an invisible event (Equation 11). We claim that the sequence of coalescent times sets $\mathbf{t}^0, \mathbf{t}^1, \ldots, \mathbf{t}^{m-1}$ and $F^0, F^1, \ldots, F^{m-1}$ matrices corresponding to the ranked tree shapes of local genealogies $g_0, g_1, \ldots, g_{m-1}$ are sufficient statistics to infer $N(t)$ under the SMC′ process. We prove this through *Propositions 3–6*.

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.

$$\begin{aligned}\Pr\left[G^T = \{F, t_n, t_{n-1}, \ldots, t_2\} | N(t)\right] &= \Pr[t_n, t_{n-1}, \ldots, t_2 | N(t)] \Pr[F | t_n, t_{n-1}, \ldots, t_2] \\ &= \frac{n!(n-1)!}{2^{n-1}} \Pr[G = \{K_n, t_n, \ldots, t_2\} | N(t)] \frac{2^{n-c-1}}{(n-1)!} \\ &= \frac{n!}{2^c} \prod_{j=2}^{n} \frac{1}{N(t_j^0)} \exp\left\{ - \int_{t_{j+1}^0}^{t_j^0} \frac{A^0(t)\left(A^0(t)-1\right)dt}{2N(t)} \right\}. \end{aligned} \tag{A7}$$

■

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_{i-1}^T to g_i^T (Figure S1A in File S1 and Figure S2D in File S4). In fact, the $I^{i-1}(t)$ function defined in the *Visible transitions* subsection (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 $j \in \{2, \ldots, n\}$, $t_j^{i-1} = t_{\text{del}}^i$ and $t_j^i = t_{\text{new}}^i$. Then

$$I^{i-1}(t) = \begin{cases} 0, & \text{if } t > \min\left(t_{\text{new}}^i, t_{\text{del}}^i\right), \\ F_{l,j}^{i-1} - F_{l,j-1}^{i-1}, & \text{if } t \in \left(t_{l+1}^{i-1}, t_l^{i-1}\right) \quad \text{for} \quad l = j, j+1, \ldots, n. \end{cases}$$

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

$$\Pr\left[G_i = \left\{K^{i-1}, \mathbf{t}^i\right\} | g_{i-1} = \left\{K^{i-1}, \mathbf{t}^{i-1}\right\}, N(t)\right] = \Pr\left[G_i^T = \left\{F^{i-1}, \mathbf{t}^i\right\} | g_{i-1}^T = \left\{F^{i-1}, \mathbf{t}^{i-1}\right\}, N(t)\right].$$

When the labeled topologies of g_{i-1} and g_i are different, but the children of t_{del}^i and the children of t_{new}^i are the same, we cannot exactly identify the pruning branch and the new coalescing branch (Figure S1B in File S1) and then a transition from g_{i-1} to g_i contains the same information about pruning location as a transition from g_{i-1}^T to g_i^T. Let $t_j^{i-1} = t_{\text{del}}^i$ and $t_k^i = t_{\text{new}}^i$; since the children of t_j^{i-1} and t_k^i are the same, it is enough to consider F^{i-1}. Then

$$I^{i-1}(t) = \begin{cases} 0, & \text{if } t > \min\left(t_{\text{new}}^i, t_{\text{del}}^i\right), \\ F_{l,j}^{i-1} - F_{l,j-1}^{i-1}, & \text{if } t \in \left(t_{l+1}^{i-1}, t_l^{i-1}\right) \quad \text{for} \quad l = j, j+1, \ldots, n \end{cases}$$

and

$$\Pr\left[G_i = \left\{K^i, \mathbf{t}^i\right\} | g_{i-1} = \left\{K^{i-1}, \mathbf{t}^{i-1}\right\}, N(t)\right] = \Pr\left[G_i^T = \left\{F^{i-1}, \mathbf{t}^i\right\} | g_{i-1}^T = \left\{F^{i-1}, \mathbf{t}^{i-1}\right\}, N(t)\right].$$

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

$$\begin{aligned}\Pr[G_i = g_i | g_{i-1}, N(t)] = &\left(\frac{1}{2}\right)^{1\left\{t_j^{i-1} = t_{\text{del}}^i\right\} 1\left\{F_{n,j}^{i-1} = F_{n,j+1}^{i-1} - 2\right\}} \times \left(\frac{1}{2}\right)^{1\left\{t_j^i = t_{\text{new}}^i\right\} 1\left\{F_{n,j}^i = F_{n,j+1}^i - 2\right\}} \\ &\times \Pr\left[G_i^T = g_i^T | g_{i-1}^T, N(t)\right]. \end{aligned}$$

■

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.

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

since all that is 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 3–5* needed to express the likelihood of partially observed embedded SMC′ chain (Equation 13).

■

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.177980/-/DC1

Bayesian Nonparametric Inference of Population Size Changes from Sequential Genealogies

Julia A. Palacios, John Wakeley, and Sohini Ramachandran

DOI: 10.1534/genetics.115.177980

Supporting Information

File S1: Visible Transitions

Figure SA shows an example of a visible transition when the coalescent topology remains the same and Figure SB shows an example of a visible transition when the coalescent topology changes (coalescence between the (a,(b,c)) branch and the (d,e) branch happens after coalescence of (f) and (g) on the left tree and before coalescence of (f) and (g) on the right tree.) 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}.

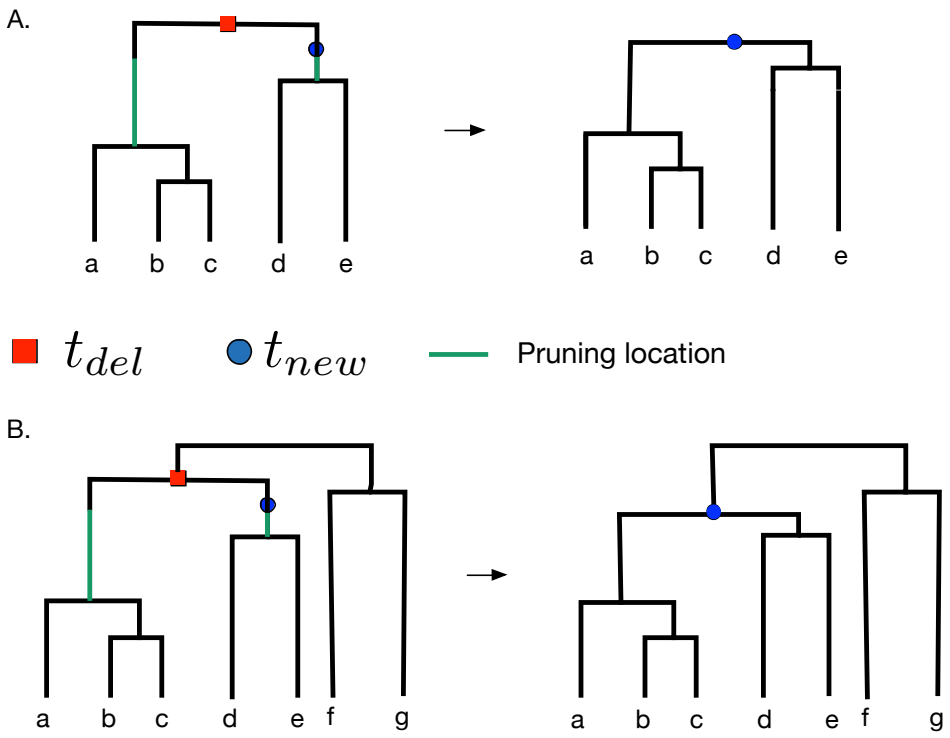


Figure S1: Examples of visible transitions when the pruning branch is uncertain. Red circle indicates deleted node at coalescent time t_{del}, blue circle indicates new node at coalescent time t_{new}. Green lines indicates possible pruning locations that could have resulted in such a visible transition.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

File S2: Visible transitions between Tajima's genealogies

A Tajima's genealogy g^T corresponds to the pair of coalescent times and a ranked tree shape with n tips (i.e. with no labels but ranked coalescent events). In Figure S, we show four possible visible transitions. In the first case (Figure SA), 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 SB, 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 SC-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.

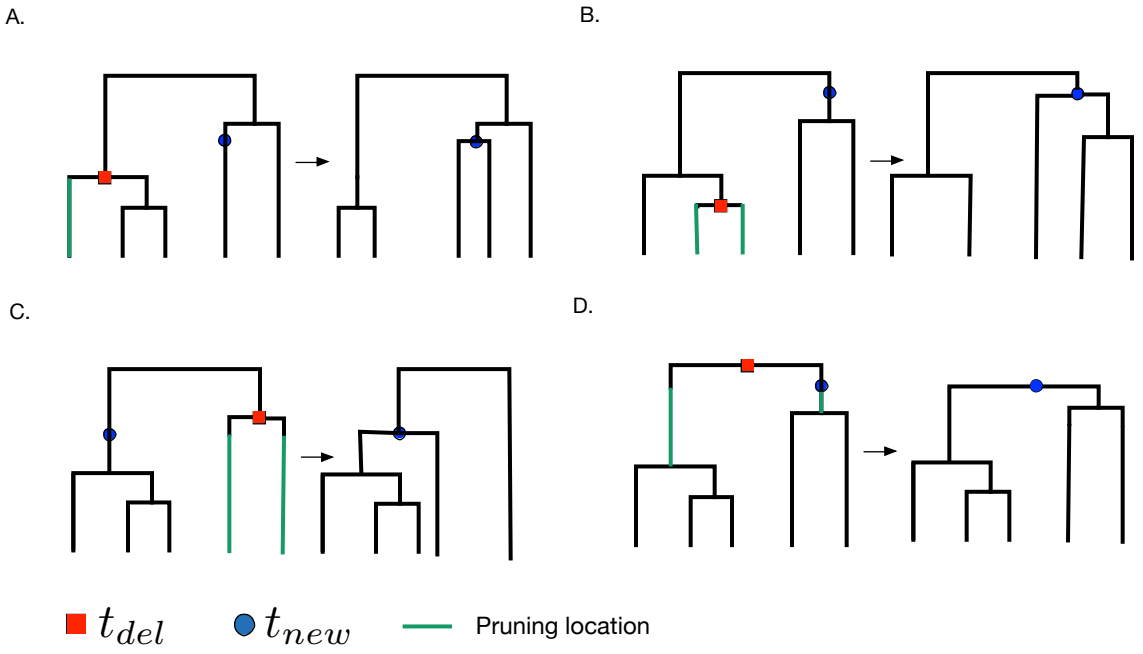


Figure S2: Examples of visible transitions between local Tajima's genealogies. Red circle indicates deleted node at coalescent time t_{del}, blue circle indicates new node at coalescent time t_{new}. Green lines indicates possible pruning locations that could have resulted in such a visible transition.

File S3: 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)

File S4: 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.

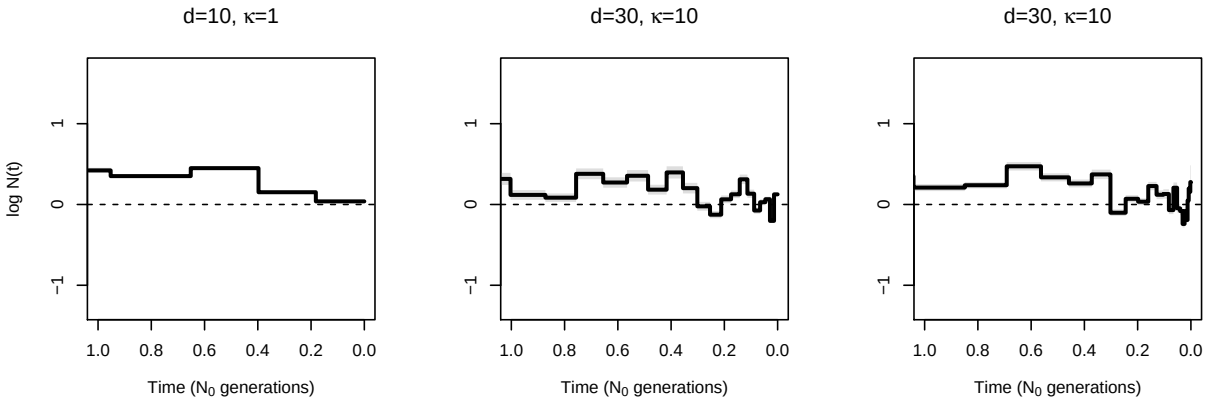


Figure S3: **EM sensitivity to parameter discretization.** Comparison of population size trajectories estimated from 1000 simulated genealogies ($m = 1000$) of 100 individuals with a constant population size. EM inference with different discretizations varying the parameters in Equation 15.

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%

File S5: 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
```

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 S4, we multiplied our estimates by $1/(8 \times 11,532)$ and converted them in logarithmic scale.

We note that *ARGweaver* assumes the SMC and not the SMC$'$ model so our estimates of $N(t)$ are biased. One source of such a bias is that the $A^i(t)$ functions that indicate the number of lineages present at time t in the SMC$'$ are replaced by $A^i(t) - 1$ if the pruning branch is present at time t in the SMC. Another source of bias is the lack of invisible recombination events in *ARGweaver* realizations. To approximate the effect of this difference we re-run our algorithm replacing $A^i(t)$ by $A^i(t) - 1$. Figure S4 shows that the main conclusions about inferred recent past population sizes remain valid; in our analysis of human data (Figure 8) we only focus on the recent past.

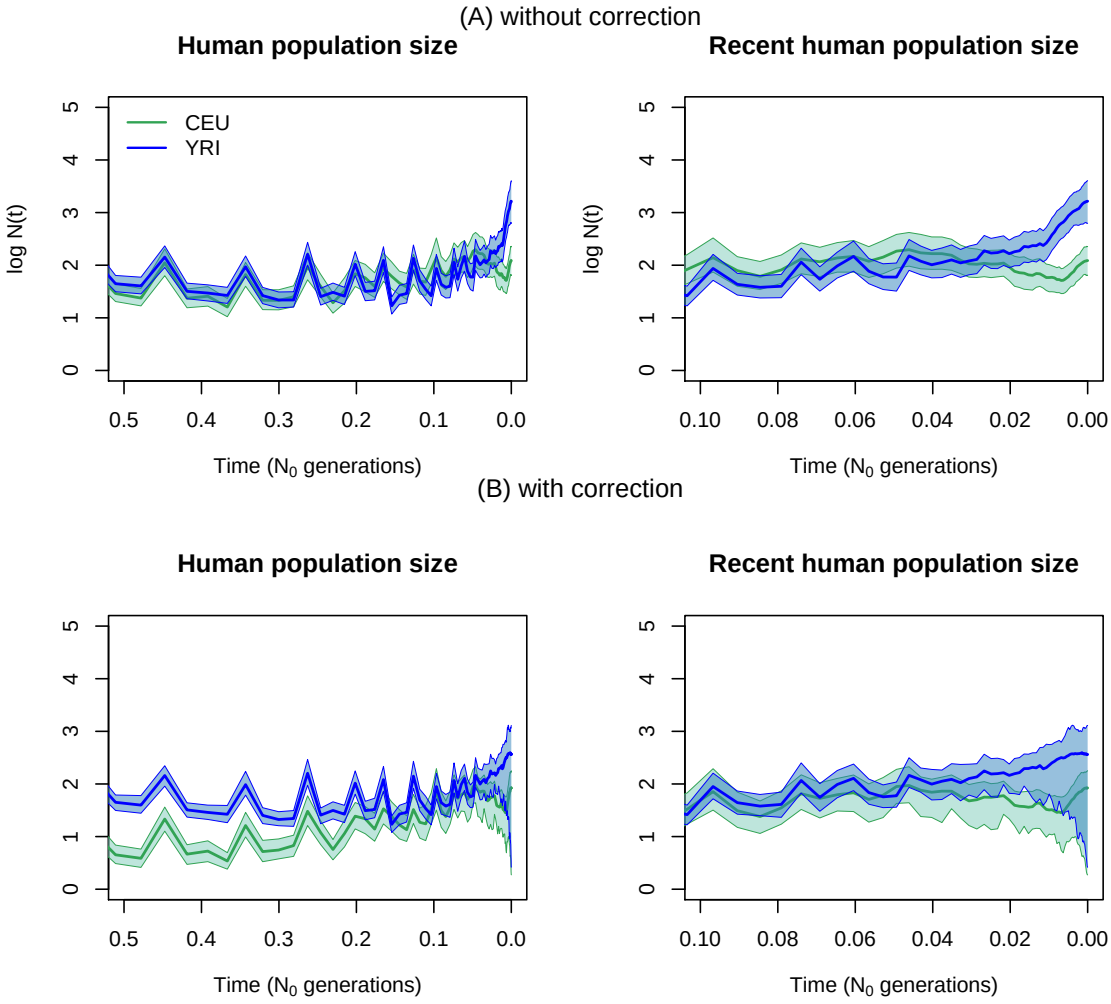


Figure S4: **Inference of human population size trajectories** $N(t)$ **for** $n = 10$**.** Green solid line and green shaded areas represent the posterior median and 95% BCI for European population (CEU) and blue solid line and blue shaded areas represent the posterior median and 95% BCI for Yoruban population. **(A) Without correction.** We ignore the fact that our genealogies were generated assuming the SMC process instead of SMC′. **(B) With correction.** We corrected the function of the number of lineages to approximate the SMC likelihood. Figures on the right show the same results as in the left side for the recent past $(0,0.1N_0)$.

File S6: Sampling more individuals on simulations

In Figure S5 we re-arrange our results on simulations shown in Figures 5−7 to compare our estimations when increasing the number of samples. We find that increasing n does not necessarily improve estimation from 1000 local genealogies.

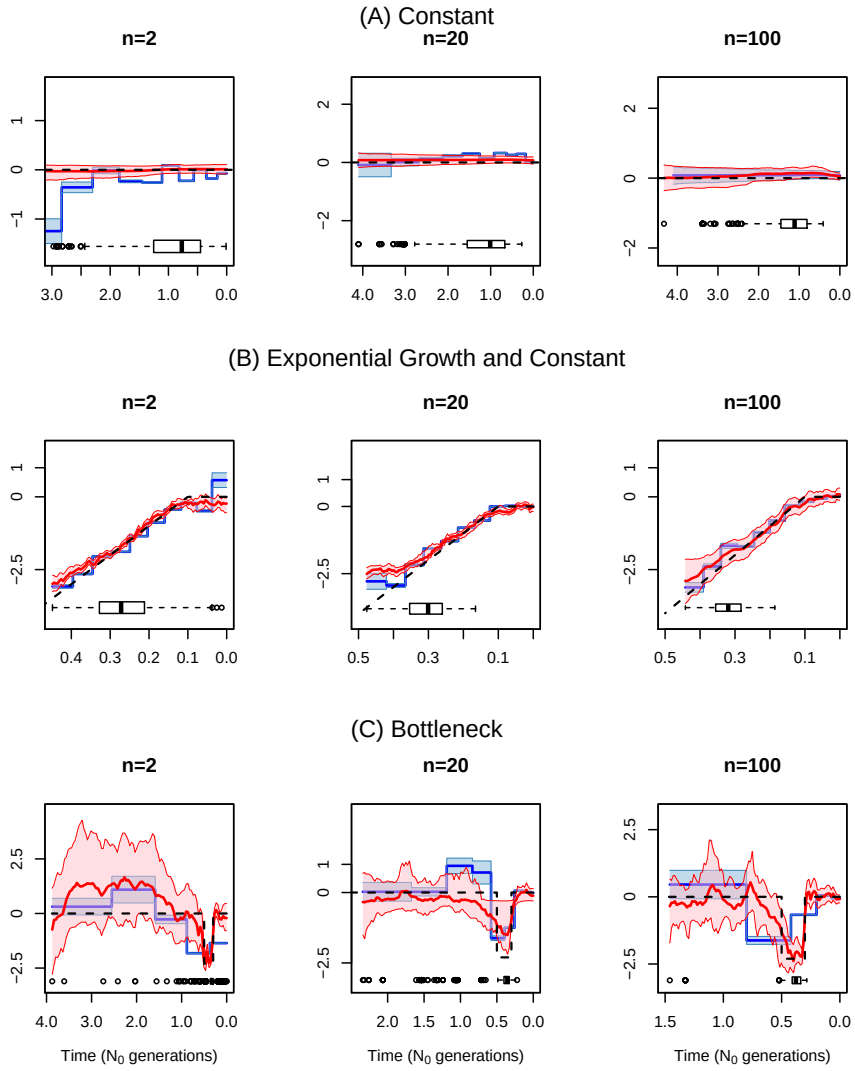


Figure S5: **Comparison of population size trajectories $N(t)$ inferred varying the number of samples from** 1000 **simulated local genealogies** (A) Simulated data under constant population size, (B) exponential and constant trajectory, and (C) a bottleneck. We rearrange the plots displayed in Figures 5−7 corresponding to 1000 simulated genealogies. Red curves are pink areas correspond to our Bayesian GP estimates and blue curves and areas correspond to the EM estimates.

File S7: Fisher Information Calculation

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

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

$$\mathrm{E}[\Delta_j^i \Delta_k^l \mid \mathbf{Y}] = \begin{cases} \Delta_{j,k}^i & i = l \\ \mathrm{E}[\Delta_j^i \mid \mathbf{Y}]\mathrm{E}[\Delta_k^l \mid \mathbf{Y}] & i \neq l \end{cases}$$

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

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

$$(z\Delta)_{jk}^i = (x_{k+1}^i - x_k^i)\frac{\sum_{l=1}^{k-1} \hat{F}_{l,j}^i \hat{P}_{l,j}^i}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1} \hat{F}_{k,j}^i \hat{P}_{k,j}^i}$$

$$+ \int_{x_k^i}^{x_{k+1}^i} (x_{k+1}^i - u) \exp\left\{-\frac{(x_{k+1}^i - u)A^i(x_{k+1}^i)}{N(x_{k+1}^i)}\right\} du \frac{\hat{F}_{k,j}^i \frac{\hat{P}_{k,j}^i}{\hat{Q}_k^i}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1} \hat{F}_{k,j}^i \hat{P}_{k,j}^i}$$

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

$$(z\Delta)_{jk}^i = z_j^i E[\Delta_k^i \mid \mathbf{Y}]$$

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

$$\Delta_{j,k}^i = (x_{j+1}^i - x_j^i)(x_{k+1}^i - x_k^i)\frac{\sum_{l=1}^{j-1}\sum_{m=k+1}^{D_i-1} \hat{F}_{l,m}^i \hat{P}_{l,m}^i}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1} \hat{F}_{k,j}^i \hat{P}_{k,j}^i}$$

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

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

and

$$\Delta_{j,j}^i = (x_{j+1}^i - x_j^i)^2 \frac{\sum_{k=1}^{j-1}\sum_{l=j+1}^{D_i-1} \hat{F}_{k,l}^i \hat{P}_{k,l}^i}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1} \hat{F}_{k,j}^i \hat{P}_{k,j}^i}$$

$$+ \int_{x_j^i}^{x_{j+1}^i} (x_{j+1}^i - u)^2 \exp\left\{-\frac{(x_{j+1}^i - u)A^i(x_{j+1}^i)}{N(x_{j+1}^i)}\right\} du \frac{\sum_{k=j+1}^{D_i-1} \hat{F}_{j,k}^i \frac{\hat{P}_{j,k}^i}{\hat{Q}_j^i}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1} \hat{F}_{k,j}^i \hat{P}_{k,j}^i}$$

$$+\int_{x_j^i}^{x_{j+1}^i}\int_u^{x_{j+1}^i}(t-u)^2\frac{1}{N(x_{j+1}^i)}\exp\left\{-\frac{(t-u)A^i(x_{j+1}^i)}{N(x_{j+1}^i)}\right\}dtdu\frac{\hat{F}_{j,j}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}_{k,j}^i\hat{P}_{k,j}^i}$$

$$+\int_{x_j^i}^{x_{j+1}^i}(t-x_j^i)^2\exp\left\{-\frac{(t-x_j^i)A^i(x_{j+1}^i)}{N(x_{j+1}^i)}\right\}dt\frac{\sum_{k=1}^{j-1}\hat{F}_{k,j}^i\frac{\hat{P}_{k,j}^i}{1-\hat{q}_j^i}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}_{k,j}^i\hat{P}_{k,j}^i}$$

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

$$\Delta_{j,k}^i=\begin{cases}0 & \sum_{l=1}^j I^i(x_{l+1}^i)=0 \text{ or } y_j^i=0\\ (x_{j+1}^i-x_j^i)(x_{k+1}^i-x_k^i) & I^i(x_{j+1}^i)=0,\sum_{l=1}^j I^i(x_{l+1}^i)>0,y_j^i=1,y_k^i=1\\ (x_{k+1}^i-x_k^i)\delta_j^i & I^i(x_{j+1}^i)=1,\sum_{l=1}^j I^i(x_{l+1}^i)>0,y_k^i=1\end{cases},$$

where δ_j^i is as defined in Equation 29, and

$$\Delta_{j,j}^i=\begin{cases}0 & \sum_{l=1}^j I^i(x_{l+1}^i)=0 \text{ or } y_j^i=0\\ (x_{j+1}^i-x_j^i)^2 & I^i(x_{j+1}^i)=0,\sum_{l=1}^j I^i(x_{l+1}^i)>0,y_j^i=1\\ \delta_{j,j}^i & I^i(x_{j+1}^i)=1,\sum_{l=1}^j I^i(x_{l+1}^i)>0,y_j^i>0\end{cases},$$

where

$$\delta_{j,j}^i=(x_{j+1}^i-x_j^i)^2\left[\frac{\sum_{k=1}^{j-1}I^i(x_{k+1}^i)\hat{Q}_k^i\prod_{l=k+1}^{D_i-1}[\hat{q}_l^i]^{y_l^i}}{\sum_{k=1}^{D_i-1}I^i(x_{k+1}^i)\hat{Q}_k^i\prod_{l=k+1}^{D_i-1}[\hat{q}_l^i]^{y_l^i}}\right]$$

$$+\int_{x_j^i}^{x_{j+1}^i}(x_{j+1}^i-u)^2\exp\left\{-\frac{(x_{j+1}^i-u)A^i(x_{j+1})}{N(x_{j+1}^i)}\right\}du\left[\frac{I^i(x_{j+1}^i)\prod_{l=j+1}^{D_i-1}[\hat{q}_l^i]^{y_l^i}}{\sum_{k=1}^{D_i-1}I^i(x_{k+1}^i)\hat{Q}_k^i\prod_{l=k+1}^{D_i-1}[\hat{q}_l^i]^{y_l^i}}\right]$$

and

For $i\in\mathcal{I}$

$$\mathrm{E}[z_j^i\Delta_j^i\mid\mathbf{Y}]=\int_{x_j^i}^{x_{j+1}^i}\int_u^{x_{j+1}^i}(t-u)\frac{1}{N(x_{j+1}^i)}\exp\left\{-\frac{(t-u)A^i(x_{j+1}^i)}{N(x_{j+1}^i)}\right\}dtdu\frac{\hat{F}_{j,j}}{\sum_{j=1}^{D_i-1}\sum_{k=j}^{D_i-1}\hat{F}_{k,j}^i\hat{P}_{k,j}^i}$$

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

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

$$\frac{\partial}{\partial\log N_l}\mathcal{L}_c(\mathbf{Y}_c;\hat{\mathbf{N}})=B_l-A_l+C_l-Z_l \tag{1}$$

With

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

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

$$C_l = \sum_{i=0}^{m-2} \sum_{j=1}^{D} A^i(x^i_{j+1}) \Delta^i_j 1^i_{l,j} \exp[-\log N_l],$$

and

$$Z_l = \sum_{i=0}^{m-2} \sum_{j=1}^{D} z^i_j 1^i_{l,j}.$$

Next, differentiating Equation 1 in this file, 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

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

$$+2 \exp[-2logN_l] \sum_{i=0}^{m-2} \sum_{j=1}^{D_i-1} \left[A^i(x^i_{j+1}) \mathrm{E}[\Delta^i_j \mid \mathbf{Y}] 1^i_{l,j} \sum_{p=i+1}^{m-2} \sum_{k=1}^{D_p-1} A^p(x^p_{k+1}) \mathrm{E}[\Delta^p_k \mid \mathbf{Y}] 1^p_{l,k} \right],$$

$$\mathrm{E}[Z_l^2 \mid \mathbf{Y}] = \sum_{i=0}^{m-2} \sum_{j=1}^{D_i-1} \left[\mathrm{E}[z^i_j \mid \mathbf{Y}] 1^i_{l,j} + 2\mathrm{E}[z^i_j \mid \mathbf{Y}] 1^i_{l,j} \sum_{p=i+1}^{m-2} \sum_{k=1}^{D_p-1} \mathrm{E}[z^p_k \mid \mathbf{Y}] 1^p_{l,k} \right]$$

and

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

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

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

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

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

and for $l < o$

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

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.