Bayesian Estimation of Population Size Changes by Sampling Tajima's Trees

Julia A. Palacios[1,2,*], Amandine Véber[3], Lorenzo Cappello[1], Zhangyuan Wang[4], John Wakeley[5] and Sohini Ramachandran[6,7]

*corresponding author: juliapr@stanford.edu

Abstract

The large state space of gene genealogies is a major hurdle for inference methods based on Kingman's coalescent. Here, we present a new Bayesian approach for inferring past population sizes which relies on a lower resolution coalescent process we refer to as "Tajima's coalescent". Tajima's coalescent has a drastically smaller state space, and hence it is a computationally more efficient model, than the standard Kingman coalescent model. We provide a new algorithm for efficient and exact likelihood calculations, which exploits a directed acyclic graph and a correspondingly tailored Markov Chain Monte Carlo method. We compare the performance of our Bayesian Estimation of population size changes by Sampling Tajima's Trees (BESTT) with a popular implementation of coalescent-based inference in BEAST using simulated data and human data. We empirically demonstrate that BESTT can accurately infer effective population sizes, and it further provides an efficient alternative to the Kingman's coalescent. The algorithms described here are implemented in the `R` package `phylodyn`, which is available for download at `https://github.com/JuliaPalacios/phylodyn`.

1 Introduction

Modeling gene genealogies from an alignment of sequences — timed and rooted bifurcating trees reflecting the ancestral relationships among sampled sequences — is a key step in coalescent-based inference of evolutionary parameters such as effective population sizes. In the neutral coalescent model without recombination, observed sequence variation is produced by a stochastic process of mutation acting along the branches of the gene genealogy (Kingman, 1982a; Watterson, 1975), which is modeled as a realization of the coalescent point process at a neutral non-recombining locus. In the coalescent point process, the rate of coalescence (the merging of two lineages into a common ancestor at some time in the past) is a function that varies with time, and it is inversely proportional to the effective population size at time t, $N(t)$ (Kingman, 1982b; Slatkin and Hudson,

[1]Department of Statistics. Stanford University, Stanford, CA 94305.
[2]Department of Biomedical Data Science. Stanford School of Medicine, Stanford, CA 94305.
[3]CMAP, École Polytechnique, CNRS, Palaiseau, France
[4]Department of Computer Science. Stanford University, Stanford, CA 94305
[5]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138
[6]Center for Computational Molecular Biology, Brown University, Providence, RI 02912
[7]Department of Ecology and Evolutionary Biology, Brown University, Providence, RI 02912

1991; Donnelly and Tavaré, 1995). Our goal is to infer $(N(t))_{t\geq 0}$ which we will refer to as the "effective population size trajectory".

Multiple methods have been developed to infer $(N(t))_{t\geq 0}$ using the standard coalescent model on genomic sequence datasets (Griffiths and Tavaré, 1996; Kuhner and Smith, 2007; Minin et al., 2008; Li and Durbin, 2011; Drummond et al., 2012; Palacios and Minin, 2013; Gill et al., 2013; Sheehan et al., 2013; Palacios et al., 2015). These methods must contend with two major challenges: (*i*) choosing a prior distribution or functional form for $(N(t))_{t\geq 0}$, and (*ii*) integrating over the large hidden state space of genealogies. For example, several previous approaches have assumed exponential growth (Griffiths and Tavaré, 1996; Kuhner et al., 1998; Kuhner and Smith, 2007), in which case the estimation of $(N(t))_{t\geq 0}$ is reduced to the estimation of one or two parameters. In general, the functional form of $(N(t))_{t\geq 0}$ is unknown and needs to be inferred. A commonly used naive nonparametric prior on $(N(t))_{t\geq 0}$ is a piecewise linear or constant function defined on time intervals of constant or varying sizes (Heled and Drummond, 2008; Sheehan et al., 2013; Schiffels and Durbin, 2014). The specification of change points in such time-discretized effective population size trajectories is inherently difficult because it can lead to runaway behavior or large uncertainty in $(\widehat{N}(t))_{t\geq 0}$ (Minin et al., 2008; Heled and Drummond, 2008; Li and Durbin, 2011; Sheehan et al., 2013; Palacios et al., 2015). These difficulties can be avoided by the use of Gaussian-process priors in a Bayesian nonparametric framework, allowing accurate and precise estimation (Palacios and Minin, 2013; Gill et al., 2013; Lan et al., 2014; Palacios et al., 2015).

The second challenge for coalescent-based inference of $(N(t))_{t\geq 0}$ is the integration over the hidden state space of genealogies. Given molecular sequence data $\mathbf{Y}$ and a mutation model with vector of parameters $\boldsymbol{\mu}$, current methods rely on calculating the marginal likelihood function $\Pr(\mathbf{Y}|(N(t))_{t\geq 0}, \boldsymbol{\mu})$ by integrating over all possible coalescence and mutation events. Under the infinite-sites mutation model without intra-locus recombination (Watterson, 1975), this integration requires a computationally expensive importance sampling technique or Markov Chain Monte Carlo (MCMC) techniques (Griffiths and Tavaré, 1994a; Stephens and Donnelly, 2000; Hobolth et al., 2008; Wu, 2010). Moreover, a maximum likelihood estimate of $(N(t))_{t\geq 0}$ cannot be explicitly obtained; instead, it is obtained by exploring a grid of parameter values. For finite-sites mutation models, current methods approximate the marginal likelihood function by integrating over all possible genealogies via MCMC methods (Equation (1); Kuhner (2006); Drummond et al. (2012)). Both cases may be represented as

$$\Pr\big(\mathbf{Y}|(N(t))_{t\geq 0}, \mu\big) = \int \Pr(\mathbf{Y}|\mathbf{g}, \mu)\Pr\big(\mathbf{g}|(N(t))_{t\geq 0}\big)d\mathbf{g}, \tag{1}$$

in which $\Pr(\cdot)$ is used to denote both the probability of discrete variables and the density of continuous variables. The integral above involves an $(n-1)$-dimensional integral over $n-1$ coalescent times and a sum over all possible tree topologies with n leaves. Therefore, these methods require a very large number of MCMC samples, and exploration of the posterior space of genealogies continues to be an active area of research (Kuhner et al., 1998; Rannala and Yang, 2003; Drummond et al., 2012; Whidden and Matsen, 2015; Aberer et al., 2016).

Current methods rely on the Kingman n-coalescent process to model the sample's ancestry. However, the state space of genealogical trees grows superexponentially with the number of samples, making inference computationally challenging for large sample sizes. In this study, we develop a Bayesian nonparametric model that relies on Tajima's coalescent, a lower resolution coalescent process with a drastically smaller state space than that of Kingman's coalescent. In Cappello and Palacios (2019), the authors quantify this striking reduction in cardinality. In particular in

this study, we infer the posterior distribution $\Pr((N(t))_{t\geq 0}, \mathbf{g}^T, \boldsymbol{\tau} \mid \mathbf{Y}, \mu)$, where $\mathbf{g}^T$ corresponds to the Tajima's genealogy of the sample (see Figure 1A and Section 2.4), $(\log N(t))_{t\geq 0}$ has Gaussian process prior with precision hyperparameter τ, and mutations occur according to the infinite-sites model of Watterson (1975). This results in a new more efficient method for inferring $(N(t))_{t\geq 0}$ called **B**ayesian **E**stimation by **S**ampling **T**ajima's **T**rees (BESTT), with a drastic reduction in the state space of genealogies. We show using simulated data that BESTT can accurately infer effective population size trajectories and that it provides a more efficient alternative than Kingman's coalescent models.

Next, we start with an overview of BESTT, detail our representation of molecular sequence data and define the Tajima coalescent process. We then introduce a new augmented representation of sequence data as a directed acyclic graph (DAG). This representation allows us to both calculate the conditional likelihood under the Tajima coalescent model, and to sample tree topologies compatible with the observed data. We then provide an algorithm for likelihood calculations and develop an MCMC approach to efficiently explore the state space of unknown parameters. Finally, we compare our method to other methods implemented in BEAST (Drummond et al., 2012) and estimate the effective population size trajectory from human mtDNA data. We close with a discussion of possible extensions and limitations of the proposed model and implementation.

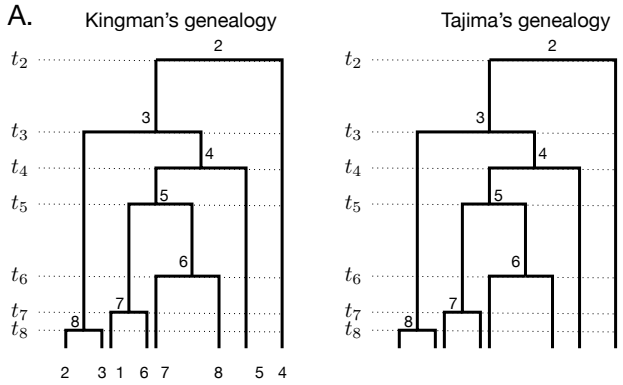
2 Methods/Theory

2.1 Overview of BESTT

Our objective in the implementation of BESTT is to estimate the posterior distribution of model parameters by replacing Kingman's genealogy with Tajima's genealogy $\mathbf{g}^T$. A Tajima's genealogy does not include labels at the tips (Figure 1): we do not order individuals in the sample but label only the lineages that are ancestral to at least two individuals (that is, we only label the internal nodes of the genealogy). Replacing Kingman's genealogy by Tajima's genealogy in our posterior distribution exponentially reduces the size of the state space of genealogies (Figure 1B). In order to compute $\Pr(\mathbf{Y}|\mathbf{g}^T, \mu)$, the conditional likelihood of the data conditioned on a Tajima's genealogy, we assume the infinite sites model of mutations and leverage a directed acyclic graph (DAG) representation of sequence data and genealogical information. Note that the overall likelihood, Eq. (1), will differ only by a combinatorial factor from the corresponding likelihood under the Kingman coalescent. Our DAG represents the data with a gene tree (Griffiths and Tavaré, 1994a), constructed via a modified version of the perfect phylogeny algorithm of Gusfield (1991). This provides an economical representation of the uncertainty and conditional independences induced by the model and the observed data.

Under the infinite-sites mutation model, there is a one-to-one correspondence between observed sequence data and the gene tree of the data (Gusfield, 1991) (Sections 2.2-2.3). We further augment the gene tree representation with the allocation of the number of observed mutations along the Tajima's genealogy to generate a DAG (Section 2.5). The conditional likelihood $\Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu)$ is then calculated via a recursive algorithm that exploits the auxiliary variables defined in the DAG nodes, marginalizing over all possible mutation allocations (Section 2.6). We approximate the joint posterior distribution $\Pr((N(t))_{t\geq 0}, \mathbf{g}^T, \boldsymbol{\tau} \mid \mathbf{Y}, \mu)$ via an MCMC algorithm using Hamiltonian Monte Carlo for sampling the continuous parameters of the model and a novel Metropolis-Hastings algorithm for sampling the discrete tree space.

A.



B.

n	labeled histories	ranked tree shapes
3	3	1
5	180	5
7	56700	61
10	2.5×10^9	7936
12	9.3×10^{12}	353792

Figure 1: **For a sample of size n, the number of Tajima's genealogies is superexponentially fewer compared to the number of Kingman's genealogies.** **A**: A Kingman's genealogy and a Tajima's genealogy for $n = 8$. A Kingman's genealogy (left) comprises a vector of coalescent times and the labeled topology; the number of possible labeled topologies for a sample of size n is $n!(n-1)!/2^{n-1}$. A Tajima's genealogy (right) comprises a vector of coalescent times and a ranked tree shape. In both cases, coalescent events are ranked from 2 at time t_2 to n at time t_n. Coalescent times are measured from the present (time 0) back into the past. **B**. The numbers of labeled topologies and ranked tree shapes for different values of the sample size, n.

2.2 Summarizing sequence data Y as haplotypes and mutation groups

Let the data consist of n fully linked haploid sequences or alignments of nucleotides at s segregating sites sampled from n individuals at time $t = 0$ (the present). Note that any labels we afix to the individuals are arbitrary in the sense that they will not enter into the calculation of the likelihood. We further assume the infinite sites mutation model of Watterson (1975) with mutation parameter μ and known ancestral states for each of the sites. Then we can encode the data into a binary matrix $\mathbf{Y}$ of n rows and s columns with elements $y_{i,j} \in \{0, 1\}$, where 0 indicates the ancestral allele.

In order to calculate the Tajima's conditional likelihood $\Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu)$, we first record each haplotype's frequency and group repeated columns to form *mutation groups*; a mutation group corresponds to a shared set of mutations in a subset of the sampled individuals. We record the cardinality of each mutation group (*i.e.*, the number of columns grouped to form each mutation group). In Figure 2A, there are two columns labeled "b", corresponding to two segregating sites which have the exact same pattern of allelic states across the sample. Further, two individuals carry the derived allele of mutation group "b", so in this case the frequency of haplotype 7 and the cardinality mutation group "b" are both equal to 2. We denote the number of haplotypes in the sample as h, the number of mutation groups as m, and the representation of $\mathbf{Y}$ as haplotypes and mutation groups as $\mathbf{Y}_{h\times m}$.

2.3 Representing $\mathbf{Y}_{h\times m}$ as a gene tree

$\mathbf{Y}_{h\times m}$ (Figure 2A, previous section) can alternatively be represented as a gene tree (or perfect phylogeny; Gusfield (1991); Griffiths and Tavaré (1994b)). This representation relies on our assumption of the infinite sites mutation model in which, if a site mutates once in a given lineage, all descendants of that lineage also have the mutation *and no other individuals carry that mutation.* The haplotype data summarized in Figure 2A corresponds to the gene tree (perfect phylogeny) given in Figure 2B.

A *gene tree* for a matrix $\mathbf{Y}_{h\times m}$ of h haplotypes and m mutation groups is a rooted tree $\mathcal{T}$ with h leaves such that (Figure 2B):

1. Each row of $\mathbf{Y}_{h\times m}$ corresponds to exactly one leaf of $\mathcal{T}$. The black numbers at leaf nodes in Figure 2B are the haplotype frequencies.

2. Each mutation group m of $\mathbf{Y}_{h\times m}$ is represented by exactly one edge of $\mathcal{T}$. The red numbers along edges in Figure 2B give the cardinality of each mutation group (*i.e.* the number of segregating sites in each mutation group; see Figure 2A). The edges of $\mathcal{T}$ corresponding to each mutation group are labeled accordingly (letters in Figures 2A and 2B). Note that external edges need not be labeled if there are no mutations exclusively present in the group of individuals descending from those edges.

3. The labels and the numbers associated with the edges along the unique path from the root to a leaf exactly specify a row of $\mathbf{Y}_{h\times m}$.

Dan Gusfield's perfect phylogeny algorithm (Gusfield, 1991) transforms the sequence data $\mathbf{Y}_{h\times m}$ into a gene tree and this transformation is one-to-one.

2.4 Tajima's genealogies

BESTT explores the state space of Tajima's genealogies ($\mathbf{g}^T$) as opposed to Kingman's genealogies ($\mathbf{g}$) and calculates the conditional likelihood $\Pr(\mathbf{Y}|\mathbf{g}^T, \mu)$. Kingman's n-coalescent is a continuous-time Markov chain taking its values in the set of partitions of the label set $\{1, \ldots, n\}$, which we denote as $\mathcal{P}_n$. At time $t = 0$, the process starts at $\{\{1\}, \ldots, \{n\}\}$, when there are n labeled lineages, and it stops at $\{\{1, \ldots, n\}\}$, when there is a single lineage ancestral to all n individuals. A transition in the n-coalescent consists of a merger, or coalescence, of two lineages (*i.e.*, two blocks of the partition are chosen uniformly at random to merge in a single block). Sainudiin et al. (2015) describe six different resolutions of the discrete coalescent process for n lineages; these different resolutions are Markovian lumpings of the states of the coalescent process. For example, the pure death process that tracks the number of lineages over time (Kingman, 1982a) is the coarsest resolution, while the partition-valued n-coalescent provides one of the highest resolutions.

The process that keeps track of the blocks formed at each time step (including the labels of the individuals of the sample within each block, which share a common ancestor) induces a ranked labeled rooted binary tree that we call a "labeled topology". Note that when the labeled tree topology is presented together with the coalescent times, ranking of coalescent events is redundant. For this reason, only four of the six resolutions presented in Sainudiin et al. (2015) are distinguishable.

A *Kingman's genealogy* is a pair $\mathbf{g} = \{K_n, \mathbf{t}\}$, consisting of a labeled tree topology K_n, and a vector of coalescent times $\mathbf{t} = (t_n, \ldots, t_2)$ whose state space is $\mathcal{G} = \mathcal{K}_n \times \mathbb{R}_+^{n-1}$. In this study, we assume that all sequences are sampled at the same time (*i.e.* the present, or time $t = 0$) and that the coalescent times, $t_n, \ldots, t_2$, are measured from the present back into the past. In addition, we define t_k to be the time of the coalescent event which decreases the number of ancestral lineages from k to $k-1$. Thus, $t_k - t_{k+1}$ is the length of time in the ancestry of the sample during which there are exactly k lineages. The number of possible labeled tree topologies for a sample of size n is $n!(n-1)!/2^{n-1}$, and each of these is equally likely. The density of a Kingman's genealogy with effective population size trajectory $(N(t))_{t\geq 0}$, denoted by $\Pr(\mathbf{g} \mid (N(t))_{t\geq 0})$, can be factored as the product of the probability of Kingman's genealogy and the coalescent times density:

$$\Pr(\mathbf{g} \mid (N(t))_{t\geq 0}) = \Pr(K_n)\Pr(\mathbf{t} \mid (N(t))_{t\geq 0}),$$

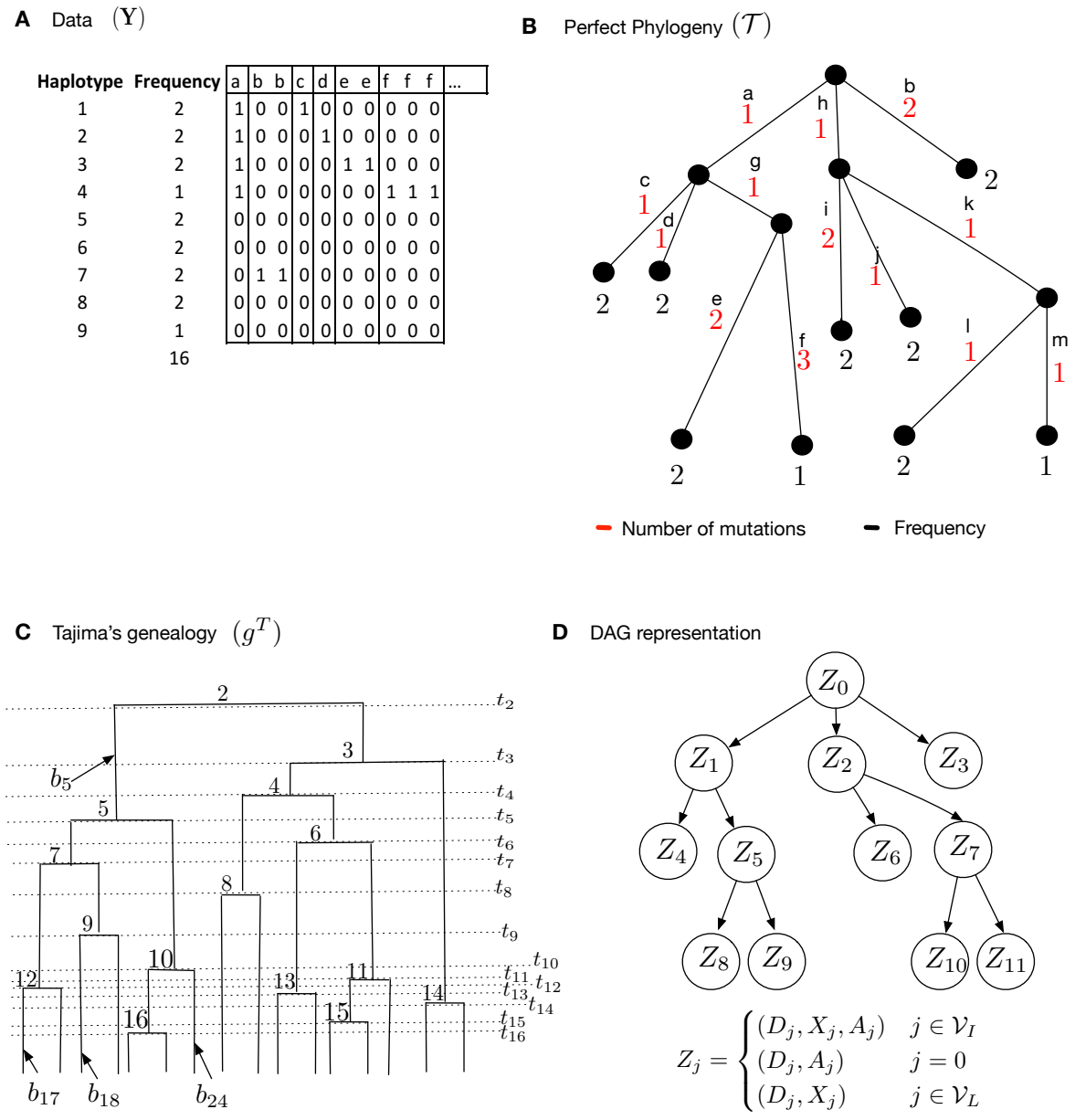


Figure 2: **Data structures exploited by our method, BESTT, for calculating the conditional likelihood of the data. A.** Compressed data representation $\mathbf{Y}_{h\times m}$ of $n = 16$ sequences and $s = 18$ (columns, only the first 10 of which are shown), comprised of 9 haplotypes and 13 mutation groups. Rows correspond to haplotypes and each polymorphic site is labeled by its mutation group $\{a, b, c, ..., m\}$. **B.** Gene tree representation of the data in panel A. Red numbers indicate the cardinality of each mutation group (number of columns with the same label in panel A). Black letters indicate the mutation group (column labels in panel A), and black numbers indicate the frequency of the corresponding haplotype. **C.** A Tajima's genealogy compatible with the gene tree in panel B. Internal nodes are labeled according to order of coalescent events from the root to the tips. Coalescent event i happens at time t_i and branches are labeled b_i (see section 2.5 for details). **D**. A Directed Acyclic Graph (DAG) representation of the gene tree in panel B together with allocation of mutation groups along the branches of the Tajima's genealogy in panel C. $\mathcal{V}_I$ denotes the set of internal nodes and $\mathcal{V}_L$ the set of leaf nodes. A detailed description of the DAG is in section 2.5.

where $\Pr(K_n) = 2^{n-1}/[n!(n-1)!]$. We refer to the distribution of a genealogy as density, although the distribution is a mixed distribution of continuous random variables (coalescent times) and discrete random variables (topology). Here again, we use the notation $\Pr(\cdot)$ to denote the probability or the density of the random variable of interest, be it continuous or discrete.

In contrast, *Tajima's genealogy* is a pair $\mathbf{g}^T = \{\mathbf{F}_n, \mathbf{t}\}$ of a ranked tree shape $\mathbf{F}_n$, and a vector of coalescent times $\mathbf{t}$ (Figure 2; Sainudiin et al. (2015); Palacios et al. (2015)). A ranked tree shape is a bifurcating tree with internal nodes labeled by their rankings (*i.e.* their order in time from past to present) and leaf labels omitted. In matrix notation, the ranked tree shape $\mathbf{F}_n$ is encoded by a triangular matrix of size $n \times n$ (Figure 3). During the interval (t_{i+1}, t_i), there are exactly i lineages ($i = 2, \ldots, n$ and by convention we set $t_{n+1} = 0$). The number of lineages through time is encoded on the diagonal of $\mathbf{F}$: $\mathbf{F}_{i,i} = i$ for i in $\{2, 3, \ldots, n\}$. For $j < i$, the entry $\mathbf{F}_{i,j}$ denotes the number of lineages that do not coalesce in the time interval (t_{i+1}, t_j); in particular, $\mathbf{F}_{i,1} = 0$ and for every i in $\{2, 3, \ldots, n\}$, $\mathbf{F}_{n,i}$ denotes the number of singletons (*i.e.*, external branches that have not coalesced) in the time interval (t_{i+1}, t_i) (Figure 3). The number of possible ranked tree shapes for a sample of size n (also called unlabeled histories, evolutionary relationships or vintaged and sized coalescent; see Sainudiin et al. (2015)) corresponds to the n-th term of the sequence A000111 of Euler zig-zag numbers (Disanto and Wiehe, 2013). The density of a Tajima's genealogy with effective population size trajectory $(N(t))_{t\geq 0}$, denoted by $\Pr(\mathbf{g}^T \mid (N(t))_{t\geq 0})$, can be factored as the product of the probability of Tajima's genealogy and the coalescent times density:

$$\Pr(\mathbf{g}^T \mid (N(t))_{t\geq 0}) = \Pr(\mathbf{F}_n)\Pr(\mathbf{t} \mid (N(t))_{t\geq 0})$$

with

$$\Pr(\mathbf{F}_n) = \frac{2^{n-c-1}}{(n-1)!}, \tag{2}$$

where c is the number of cherries (nodes with two leaves; $c = 3$ in Figure 3A), which can be expressed in terms of the entries of the matrix $\mathbf{F}_n$. Equation (2) was derived independently by both Sainudiin et al. (2015) and Palacios et al. (2015).

Observe that the density of Kingman's and Tajima's genealogies differ solely by the discrete probability corresponding to the tree topology. Using either Kingman's or Tajima's genealogies, the distribution of coalescent times can be viewed as the distribution of a point process of coalescent events at times $\mathbf{t} := (t_n, t_{n-1}, \ldots, t_2)$, where t_k indicates the time, measured from the present time 0 back into the past, when two of the k extant lineages reach a common ancestor and merge. The rate at which pairs of lineages coalesce depends on the effective population size trajectory $(N(t))_{t\geq 0}$. In contrast to the case of constant effective population size, the coalescent intervals $t_k - t_{k+1}$ for $k = 2, \ldots, n$ are not independent of each other when $N(t)$ varies with time. However, the density of a realization of the coalescent point process can be decomposed into a product of conditional densities as follows:

$$\Pr\big(\mathbf{t} \mid (N(t))_{t\geq 0}\big) = \prod_{k=2}^{n} \Pr\big(t_k - t_{k+1} \mid t_{k+1}, (N(t))_{t\geq 0}\big), \tag{3}$$

where again we set $t_{n+1} = 0$. The conditional density of the coalescent interval $t_k - t_{k+1}$ takes the following form:

$$\Pr\big(t_k - t_{k+1} | t_{k+1}, (N(t))_{t\geq 0}\big) = \frac{C_k}{N(t_k)} \exp\left[-\int_{t_{k+1}}^{t_k} \frac{C_k}{N(t)}\, dt\right] \tag{4}$$

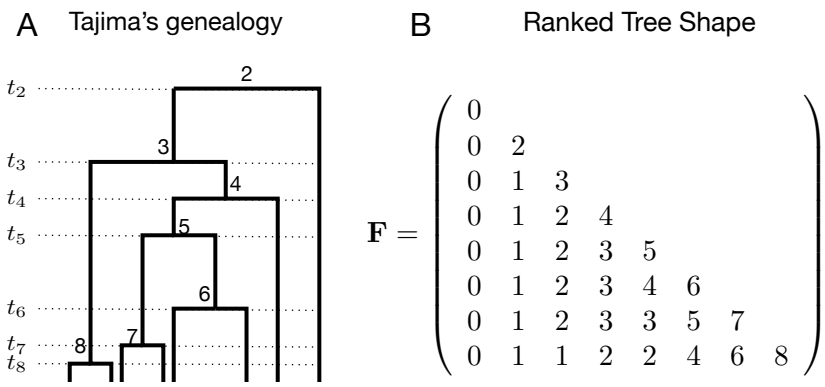


Figure 3: **Ranked tree shape** Left: Example of a Tajima's genealogy (redrawn from Figure 1A) with coalescent events ranked from 2 at time t_2 to n at time t_n. Right: The corresponding $\mathbf{F}_n$ matrix, with $n = 8$, that encodes the ranked tree shape information of the Tajima's genealogy on the left. $\mathbf{F}_{i,j}$ denotes the number of lineages that do not coalesce in the time interval (t_{i+1}, t_j). In particular, $\mathbf{F}_{n,i}$ for i in $\{2, 3, \ldots, n\}$ denotes the number of singletons (external branches that have not coalesced) in the time interval (t_{i+1}, t_i).

(Slatkin and Hudson, 1991) where $C_k = \binom{k}{2}$ is the number of possible coalescent events when k ancestral lineages are present.

2.5 An augmented data representation using directed acyclic graphs

A key component of BESTT is the calculation of the conditional likelihood $\Pr(\mathbf{Y}|\mathbf{g}^T, \mu)$. We compute the conditional likelihood recursively over a directed acyclic graph (DAG) $\mathbf{D}$. Our DAG exploits the gene tree representation $\mathcal{T}$ of the data (Figure 2B), incorporates the branch length information of the Tajima's genealogy $\mathbf{g}^T$(Figure 2C) and facilitates the recursive allocation of mutations to the branches of $\mathbf{g}^T$. Here we detail the construction of the DAG.

We construct the DAG using three pieces of information: the gene tree $\mathcal{T}$, a Tajima's genealogy $\mathbf{g}^T$ and an "allocation" of mutations along the branches of the Tajima's genealogy (Figure 2). An allocation refers to a possible mapping (compatible with the data) of the observed numbers of mutations (red numbers in Figure 2B) to branches in the Tajima's genealogy. Figure 4A shows one possible mapping for the Tajima's genealogy in Figure 2C; usually this mapping is not unique. Our construction of $\mathbf{D}$ enables an efficient recursive consideration of all possible allocations of mutations along $\mathbf{g}^T$ when computing the conditional likelihood $\Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu)$.

Constructing the DAG D. Our DAG $\mathbf{D} = \{\mathbf{Z}, E\}$ (Figure 2D) with nodes $\mathbf{Z}$ and edges E is constructed from a gene tree $\mathcal{T}$. The number of internal nodes in the DAG $\mathbf{D}$ is the same as the number of internal nodes in $\mathcal{T}$. However, sister leaf nodes in $\mathcal{T}$ with the same number of descendants are grouped together in $\mathbf{D}$. For example, the leaves in Figure 2B subtending from edges i and j are grouped into Z_6 in Figure 2D, as they both have haplotype frequency 2. However, the leaves subtending from the e and f edges are not grouped (and correspond to Z_8 and Z_9 in the DAG Figure 2D) since they have respective haplotype frequencies 2 and 1. We label the root node of $\mathbf{D}$ as Z_0 and increase the index i of each node Z_i from top to bottom, moving left to right. For $i < j$, we assign a directed edge $E_{i,j}$ if the node in $\mathcal{T}$ corresponding to Z_i is connected to the node

in $\mathcal{T}$ corresponding to Z_j. The index set of internal nodes in $\mathbf{D}$ is denoted by $\mathcal{V}_I$ and the index set of leaf nodes is denoted by $\mathcal{V}_L$.

Random variables in D. Each node in $\mathbf{D}$ represents a random vector, Z_j, which includes number of descendants, number of mutations and allocation of mutations. Although the number of descendants and number of mutations are part of the observed data rather than random variables, for ease of exposition, we use capital letters to denote all three types of information. We define the random vector Z_j as follows:

$$Z_j = \begin{cases} (D_j, X_j, A_j) & j \in \mathcal{V}_I, \\ (D_j, A_j) & j = 0 \text{ (the root node)}, \\ (D_j, X_j) & j \in \mathcal{V}_L, \end{cases}$$

where D_j denotes the number of descendants of (*i.e.*, of sampled sequences subtended by) node Z_j, X_j denotes the number of mutations separating Z_j from its parent node, and A_j denotes the allocation of mutations along $\mathbf{g}^T$ (described in detail below). The number of descendants D_j is thus the number of individuals/sequences descending from node Z_j (this information is part of $\mathcal{T}$). For internal nodes, X_j records the cardinality of a mutation group, represented as a red number along the edge $E_{i,j}$ of $\mathcal{T}$ in Figure 2B, where i is the index of the parent node of Z_j. Leaf nodes in $\mathbf{D}$ may correspond to more than one leaf node in $\mathcal{T}$, namely any sister nodes with the same number of descendants. In this case, X_j is a vector with the cardinalities of the corresponding mutation groups (see for example node Z_6 in Figure 4B).

Allocation of mutation groups along $\mathbf{g}^T$. The allocation random variables $\{A_j\}$ are constrained by the information in the Tajima's genealogy $\mathbf{g}^T$. In a given $\mathbf{g}^T$, every subtree is labeled by its ranking from past to present (Figure 3). Subtree i is subtended by branch b_i with length l_i, for $i = 2, \ldots, n$. We will assume that l_2, the length of the root branch, is 0. Let c be the number of cherries (nodes with two leaves) in $\mathbf{g}^T$; the two branches of a given cherry share the same label $b_j \in \{b_{n+1}, \ldots, b_{n+c}\}$. The actual label of external branches is arbitrary but, for ease of exposition, we first label the cherries' branches from left to right by $\{b_{n+1}, \ldots, b_{n+c}\}$; singleton branches are labeled from left to right by $b_{n+c+1}, \ldots, b_{2n-c}$ (Figure 2C). The allocation variables $\{A_j\}$ determine a possible correspondence between subtrees in $\mathbf{g}^T$ and nodes in $\mathbf{D}$: in particular, A_j indicates the branches in $\mathbf{g}^T$ that subtend the subtrees corresponding to nodes $\{Z_k\}$ if $\{Z_k\}$ are child nodes of Z_j.

Allocations of mutations to branches are usually not unique and computation of the conditional likelihood $\Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu)$ requires summing over all possible allocations. In Figure 4A we show one such possible allocation of the mutation groups of the gene tree in Figure 2B along the Tajima's genealogy in Figure 2C. For example, mutation group "a" in Figure 2B with cardinality 1 (number in red) is a mutation observed in 7 individuals (sum of black numbers of leaves descending from edge marked a). This same mutation group, "a", is shown as a red number 1 in Figure 4A allocated to branch b_5. If Z_j is an internal node, the number of mutations X_j separating it from its parent node is a vector of length 1. If Z_j is a leaf node, X_j can be a vector of length greater than 1. The length of X_j is the number of the corresponding sister nodes in $\mathcal{T}$ that were grouped together in forming node Z_j. $A_j = (A_{j,1}, \ldots, A_{j,|ch(j)|})$ denotes a collection of $|ch(j)|$ vectors of branch labels in $\mathbf{g}^T$ subtending the child-node subtrees of node Z_j. $A_{j,1}$ corresponds to the branch subtending

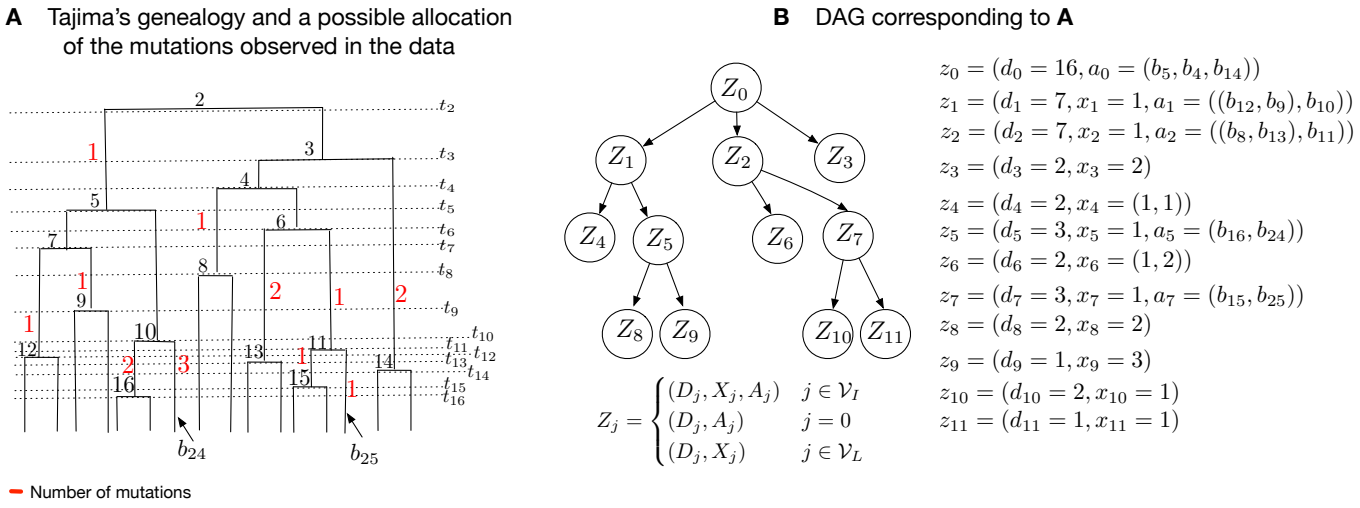


Figure 4: **DAG Construction.** **A.** A Tajima's genealogy from Figure 2C with added allocation of mutations shown in red. **B**. The corresponding augmented DAG with allocation of mutations. At the root Z_0, there are no mutations by convention. Node Z_0 has 16 descendants across 3 subtrees of $7, 7$ and 2 descendants, corresponding to nodes Z_1, Z_2, Z_3. These three subtrees subtend from b_5, b_4 and b_{14}, respectively, in $\mathbf{g}^T$ (Figure 4A). Node Z_1 corresponds to the tree subtending from b_5 of size 7 with $X_1 = 1$ mutation along b_5 and subtends three subtrees from (b_{12}, b_9) and b_{10}. Subtrees subtending from (b_{12}, b_9) are grouped together in leaf node Z_4 because they both have 2 descendants and have the same parent node. When leaf nodes represent more than one tree, such as Z_4 in Figure 4B, the random variable X_j is the vector $X_j = (X_{j,1}, X_{j,2}, \ldots, X_{j,s_j})$ that denotes the number of mutations along the branches that subtends from the tree node j that have D_j descendants.

from the leftmost child node of Z_j on $\mathbf{D}$, $A_{j,2}$ corresponds to the branch subtending from the next child node of Z_j, etc., and $A_{j,|ch(j)|}$ corresponds to the branch subtending from the rightmost child node of Z_j on $\mathbf{D}$. Observe that, since we group some of the leaf nodes in $\mathcal{T}$ into a single node in $\mathbf{D}$, any $A_{j,k}$ may be a vector of branch labels; for example $A_{1,1} = (b_{12}, b_9)$ and $A_{1,2} = b_{10}$ in Figure 4B.

2.6 Computing the conditional likelihood

Under the infinite-sites mutation model, mutations are superimposed independently on the branches of $\mathbf{g}^T$ as a Poisson process with rate μ. In order to compute $\Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu) = \Pr(\mathcal{T} \mid \mathbf{g}^T, \mu)$ we marginalize over the unspecified allocation information in the directed acyclic graph $\mathbf{D}$; that is, we

sum over all possible mappings of mutations in $\mathcal{T}$ to branches in $\mathbf{g}^T$ as follows:

$$\begin{aligned}\Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu) &= \sum_{A_0} \sum_{A_1} \cdots \sum_{A_{n_I}} \Pr(\mathbf{D} \mid \mathbf{g}^T, \mu) \\ &= \sum_{A_0} \sum_{A_1} \cdots \sum_{A_{n_I}} \Pr(Z_0, \ldots, Z_{n_I+n_L} \mid \mathbf{g}^T, \mu) \\ &= \sum_{A_0} \sum_{A_1} \cdots \sum_{A_{n_I}} \prod_{i=1}^{n_I+n_L} \Pr(Z_i \mid Z_{pa(i)}, \mathbf{g}^T, \mu)\end{aligned}$$

where $n_I = |\mathcal{V}_I|$, $n_L = |\mathcal{V}_L|$, $pa(i)$ denotes the index of the parent of node i in $\mathbf{D}$ and we set $P(Z_0 \mid \mathbf{g}^T, \mu) = 1$ because it is assumed that there are no mutations above the root node and the length of the root branch $l_2 = 0$. Writing $\mathcal{L}$ for the tree length of $\mathbf{g}^T$ (*i.e.*, the sum of the lengths of all branches of $\mathbf{g}^T$) and factoring out a global factor $e^{-\mu\mathcal{L}}$ (due to the Poisson distribution of mutations across the genealogy) from each of the above products over $i \in \{1, \ldots, n_I + n_L\}$, we have

$$\begin{aligned}&\Pr(Z_i = z_i \mid z_{pa(i)}, \mathbf{g}^T, \mu) \\ &= \begin{cases} \Pr(X_i = x_i \mid a_{pa(i)} = b_j, \mathbf{g}^T, \mu) \propto (\mu l_j)^{x_i} & \text{if } |x_i| = 1, \\ \Pr\big(X_i = (x_{i1}, \ldots, x_{ik}) \mid a_{pa(i)} = (b_{j_1}, \ldots, b_{j_k}), \mathbf{g}^T, \mu\big) \propto \sum\limits_{s \in \Pi(x_i, k)} \prod_{m=1}^{k} (\mu l_{j_m})^{s_m} & \text{if } |x_i| = k > 1, \end{cases}\end{aligned}$$

where $\Pi(x_i, k)$ is the set of all permutations of $x_i = \{x_{i1}, \ldots, x_{ik}\}$ divided into m_i groups of different sizes. The number of different permutations of the k values of x_i divided into m_i groups of sizes $k_1, \ldots, k_{m_i}$ is

$$|\Pi(x_i, k)| = \frac{k!}{\prod_{j=1}^{m_i} k_j!} \tag{5}$$

For example, assume that $x_i = \{2, 2, 2, 0, 3, 3\}$ and $a_{pa(i)} = (b_3, b_4, b_5, b_6, b_7, b_8)$ with branch lengths $\{l_3, l_4, l_5, l_6, l_7, l_8\}$. In this case, $k_1 = 3$ because there will be 3 branches with 2 mutations, $k_2 = 1$ because there will be 1 branch with 0 mutations and $k_3 = 2$ because there will be 2 branches with 3 mutations. The number of permutations of $k = 6$ mutations groups divided into $m_i = 3$ groups with cardinalities $2, 0, 3$ of sizes $3, 1, 2$ is $6!/(3!1!2!) = 60$.

The conditional likelihood $\Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu)$ is calculated via a depth first search algorithm (Appendix). The algorithm marginalizes the allocations by traversing the DAG from the tips to the root. The pseudocode can be found in the Appendix.

2.7 The case of unknown ancestral states

Up to now, we have assumed that the ancestral state was known at every segregating site. The representation of the data $\mathbf{Y}$ that we use in this case records the cardinalities of each mutation group and the genealogical relations between these groups, but does not assign labels to the sequences. Hence, in the terminology of Griffiths and Tavaré (1995), our data corresponds to an *unlabeled rooted gene tree*.

When the ancestral types are not known, the data (now denoted $\mathbf{Y}^0$) may be represented as an unlabeled *unrooted* gene tree. By the remark following Equation (1) in Griffiths and Tavaré (1995), if s is the number of segregating sites, then there are at most $s + 1$ unlabeled rooted gene

trees that correspond to the unrooted gene tree of the observed data ($\mathcal{R}(Y^0)$). By the law of total probability (see also Equation (10) in Griffiths and Tavaré (1995)), the conditional likelihood of $\mathbf{Y}^0$ can be written as the sum over all compatible unlabeled rooted gene trees $Y^{(i)}$ of the probability of $Y^{(i)}$ conditionally on $\mathbf{g}^T$. That is:

$$\Pr\big(\mathbf{Y}^0 \mid \mathbf{g}^T, \mu\big) = \sum_{i=1}^{\mathcal{R}(Y^0)} \mathrm{P}\big(\mathbf{Y}^{(i)} \mid \mathbf{g}^T, \mu\big), \tag{6}$$

where each of the $\mathbf{Y}^{(i)}$ corresponds to a unique unlabeled rooted gene tree compatible with the unrooted gene tree $\mathbf{Y}^0$ and $\mathcal{R}(Y^0)$ denotes the number of those unlabeled rooted gene trees. In the following sections, we shall assume that the ancestral type at each site is known.

2.8 Bayesian inference of the effective population size trajectory

Our posterior distribution of interest is

$$\Pr(\boldsymbol{\gamma}, \mathbf{g}^T, \tau \mid \mathbf{Y}, \mu) \propto \Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu)\Pr(\mathbf{g}^T \mid \boldsymbol{\gamma})\Pr(\boldsymbol{\gamma} \mid \tau)\Pr(\tau), \tag{7}$$

where $(\log N(t))_{t\geq 0} = (\gamma(t))_{t\geq 0} \sim \mathcal{GP}(\mathbf{0}, \mathbf{C}(\tau))$ has a Gaussian process prior with mean $\mathbf{0}$ and covariance function $\mathbf{C}(\tau)$. This specification ensures $(N(t))_{t>0}$ is non-negative. In our implementation, we assume a regular geometric random walk prior, that is, $\gamma_1 = \log N(t_1^*), \ldots, \gamma_B = \log N(t_B^*)$ at B regularly spaced time points in $[0, T]$ with

$$\mathrm{Cov}[\gamma_i, \gamma_j] = \mathrm{Cov}\big[\log N(t_i^*), \log N(t_j^*)\big] = \tau \min(t_i^*, t_j^*).$$

The parameter τ is a length scale parameter that controls the degree of regularity of the random walk. We place a Gamma prior with parameters $\alpha = .01$ and $\beta = .001$ on τ, reflecting our lack of prior information about the smoothness of the logarithm of the effective population size trajectory.

We approximate the posterior distribution of model parameters via a MCMC sampling scheme. Model parameters are sampled in blocks within a random scan Metropolis-within-Gibbs framework.

To summarize the effective population size trajectory, we compute the posterior median and 95% credible intervals pointwise at each grid point in $[0, \hat{T}]$, were $\hat{T}$ is the maximum time to the most recent common ancestor sampled.

2.8.1 Metropolis-Hastings updates for ranked tree shapes

There is a large literature on local transition proposal distributions for Kingman's topologies (Kuhner et al., 1998; Rannala and Yang, 2003; Drummond et al., 2012; Whidden and Matsen, 2015; Aberer et al., 2016). In this paper, we adapted the local transition proposal of Markovtsova et al. (2000) to Tajima's topologies. We briefly describe the scheme below and provide a pseudocode algorithm in the Appendix (Algorithm 1).

Given the current state of the chain $\{\boldsymbol{\gamma}, \tau, \mathbf{g}^T\} = \{\boldsymbol{\gamma}, \tau, \mathbf{F}_n, \mathbf{t}\}$, we propose a new ranked tree shape $\mathbf{F}^*$ in two steps: (1) we first sample a coalescent interval k uniformly on $\{1, \ldots, n-2\}$. Given k, we focus solely on the coalescent event sampled and the one that follows (thus, the last coalescent event cannot be sampled). For step (2), there are two possible scenarios: either vintage k undergoes a coalescent at event at $k+1$, or it does not. In the former scenario, we choose a new pair of lineages at random to coalesce at k from the 3 lineages subtending k and $k+1$ (excluding

k), and we coalesce the remaining lineage with k at $k+1$. In the latter scenario, we invert the order of the coalescent events; that is, vintage k is relabeled $k+1$ and conversely. The transition probability $q(\mathbf{F}_n^* \mid \mathbf{F}_n)$ is given by the product of the probabilities of the two steps. The new ranked tree shape $\mathbf{F}_n^*$ is accepted with probability given by the Metropolis-Hastings ratio defined below:

$$a_{\mathbf{F}_n} = \min\left\{1\,,\, \frac{\Pr(\mathbf{Y} \mid \mathbf{F}_n^*, \mathbf{t}, \mu)\Pr(\mathbf{F}_n^*)q(\mathbf{F}_n \mid \mathbf{F}_n^*)}{\Pr(\mathbf{Y} \mid \mathbf{F}_n, \mathbf{t}, \mu)\Pr(\mathbf{F}_n)q(\mathbf{F}_n^* \mid \mathbf{F}_n)}\right\} \tag{8}$$

2.8.2 Split Hamiltonian Monte Carlo updates of $(\boldsymbol{\gamma}, \tau)$

To make efficient joint proposals of $\boldsymbol{\gamma}$ and τ, we use the Split Hamiltonian Monte Carlo method proposed by Lan et al. (2014). The target density, $\pi(\boldsymbol{\gamma}, \tau) \propto \Pr(\mathbf{t} \mid \gamma)\Pr(\boldsymbol{\gamma} \mid \tau)\Pr(\tau)$, is the same target density implemented in Karcher et al. (2017) for fixed coalescent times $\mathbf{t}$.

2.8.3 Hamiltonian Monte Carlo updates of coalescent times

Given the current state $\{\boldsymbol{\gamma}, \tau, \mathbf{g}^T\} = \{\boldsymbol{\gamma}, \mathbf{F}_n, \mathbf{t}, \tau\}$, we propose a new vector of coalescent times with target density $\pi(\mathbf{t}') \propto P(\mathbf{Y} \mid \mathbf{F}_n, \mathbf{t}', \mu)P(\mathbf{t}' \mid \boldsymbol{\gamma})$ by numerically simulating a Hamilton system with Hamiltonian

$$H(\log(\mathbf{t}'), \mathbf{s}) = -\log\left(\pi(\log(\mathbf{t}'))\right) + \frac{1}{2}\mathbf{s}^T\mathbf{M}\mathbf{s}, \tag{9}$$

where $\mathbf{s}$ is the momentum vector assumed to be normally distributed. For our implementation, we set the mass matrix $\mathbf{M} = \mathbf{I}$, the identity matrix. We simulate the Hamiltonian dynamics of the logarithm of times to avoid proposals with negative values. Solving the equations of the Hamilton system requires calculating the gradient of the logarithm of the target density with respect to the vector of log coalescent times. The gradient of the log conditional likelihood (score function) is calculated at every marginalization step in the sum-product algorithm for the likelihood calculation.

At the beginning of Section 2.8, we described how we assume a regular geometric random walk prior on $(N(t))_{t\geq 0}$ at B regularly spaced time points in $[0, T]$. Ideally, the window size T must be at least t_2, the time to the most recent common ancestor (TMRCA). However, t_2 is not known. Our initial values of coalescent times $\mathbf{t}$ are obtained from the UPGMA implementation in `phangorn` (Schliep, 2011) with times properly rescaled by the mutation rate, and we set $T = t_2$. We initially discretize the time interval $[0, T]$ into B intervals of length $T/(B-1)$. As we generate new samples of $\mathbf{t}$, we expand or contract our grid according to the current value of t_2 by keeping the grid interval length fixed to $T/(B-1)$, effectively increasing or decreasing the dimension of $\boldsymbol{\gamma}$.

2.8.4 Local updates of coalescent times

In addition to HMC updates of coalescent times, we propose a move of a single coalescent time (excluding the TMRCA t_2) chosen uniformly at random and sampled uniformly in the intercoalescent interval; that is, we choose $i \sim U(\{n, n-1, \ldots, 3\})$ and $t_i^* \sim U(t_{i+1}, t_{i-1})$. This is a symmetric proposal and the corresponding Metropolis-Hastings acceptance probability is

$$a_{t^*} = \min\left\{1\,,\, \frac{\Pr(\mathbf{Y} \mid \mathbf{F}_n, \mathbf{t}^*, \mu)\Pr(\mathbf{t}^* \mid \boldsymbol{\gamma})}{\Pr(\mathbf{Y} \mid \mathbf{F}_n, \mathbf{t}, \mu)\Pr(\mathbf{t} \mid \boldsymbol{\gamma})}\right\}. \tag{10}$$

2.8.5 Multiple Independent loci

Thus far, we have assumed our data consist of a single linked locus of s segregating sites. We can extend our methodology to l independent loci with s_i segregating sites for $i = 1, \ldots, l$. In this case, our data $\vec{\mathbf{Y}} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_l)$ consist of l aligned sequences with elements $\{0, 1\}$, where 0 indicates the ancestral allele as before. We then jointly estimate the Tajima's genealogies $\{\mathbf{g}_i^T\}_{i=1}^l$, precision parameter τ, and vector of log effective population sizes $\boldsymbol{\gamma}$ through their posterior distribution:

$$\Pr(\boldsymbol{\gamma}, \{\mathbf{g}_i^T\}_{i=1}^l, \tau \mid \vec{\mathbf{Y}}, \boldsymbol{\mu}) \propto \left\{ \prod_{i=1}^{l} \Pr(\mathbf{Y}_i \mid \mathbf{g}_i^T, \mu_i) \Pr(\mathbf{g}_i^T \mid \boldsymbol{\gamma}) \right\} \Pr(\boldsymbol{\gamma} \mid \tau) \Pr(\tau). \tag{11}$$

In Equation (11), we enforce that all loci follow the same effective population size trajectory but every locus can have its own mutation rate μ_i.

3 Results

3.1 The performance of BESTT in applications to simulated data

We tested our new method, BESTT, on simulated data under two different demographic scenarios. Note that in this section, $N(t)$ is rescaled to the coalescent time scale, meaning that $1/N(t)$ is the pairwise rate of coalescence at time t in the past relative to the rate at the present time zero. We simulated genealogies under four different population size trajectories:

1. A period of exponential growth followed by constant size:

$$N(t) = \begin{cases} 1 & \text{if } t \in [0, 0.1), \\ \exp(1 - 10t) & \text{if } t \in [0.1, \infty). \end{cases} \tag{12}$$

2. A trajector with instantaneous growth:

$$N(t) = \begin{cases} 1 & \text{if } t \in [0, 0.05), \\ 0.05 & \text{if } t \in [0.05, \infty). \end{cases} \tag{13}$$

3. An exponential growth: $N(t) = 25e^{-5t}$

4. A constant trajectory: $N(t) = 1$

Given a genealogy of length $L = \sum_{j=2}^{n} j(t_j - t_{j+1})$, where $t_j - t_{j+1}$ is the intercoalescent length while there are j lineages, we drew the total number of mutations (segregating sites) s according to a Poisson distribution with parameter μL. We then placed the mutations uniformly at random along the branches of the genealogy. For each of the s mutations, we assigned the mutant type to individuals descending from the branch where the mutation occurred and the ancestral type otherwise.

We summarize our posterior inference $\hat{N}(t)$ by the posterior median and 95% Bayesian credible intervals after 200 thousand iterations and thinned every 10 iterations with 100 iterations of burn in. Our initial number of change points for $N(t)$ was set to 50 over the time interval between 0 and the initial time to the most recent common ancestor t_2 for all simulations; however, over the course

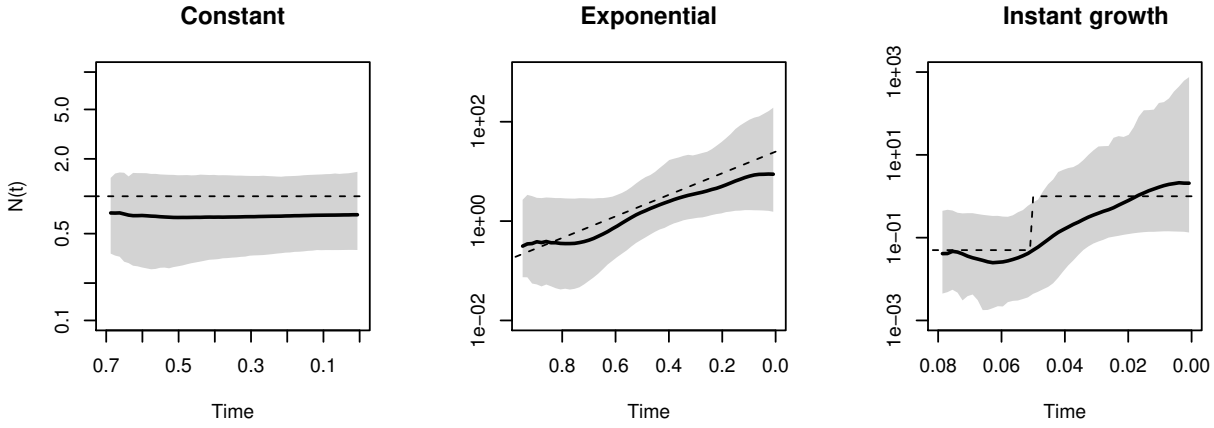


Figure 5: **Different trajectories.** Posterior inference from simulated data of $n = 10$ sequences under constant, exponential and instantaneous growth population size trajectories (dashed lines). Posterior medians are depicted as solid black lines and 95% Bayesian credible intervals are depicted by shaded areas.

of MCMC iterations, this number could increase or decrease according to the posterior distribution of t_2.

We assess accuracy and precision of our estimates using the sum of relative errors (SRE)

$$SRE = \sum_{i=1}^{k} \frac{|\hat{N}(\omega_i) - N(\omega_i)|}{N(\omega_i)}, \tag{14}$$

where $\hat{N}(\omega_i)$ is the estimated effective population size trajectory at time ω_i. Second, we computed the mean relative width as

$$MRW = \sum_{i=1}^{k} \frac{|\hat{N}_{up}(\omega_i) - \hat{N}_{lo}(\omega_i)|}{kN(\omega_i)}, \tag{15}$$

where $\hat{N}_{up}(\omega_i)$ corresponds to the 97.5% upper limit and $\hat{N}_{lo}(\omega_i)$ corresponds to the 2.5% lower limit of the estimated posterior distribution of $N(\omega_i)$. In addition, we measured how well the 95% credible intervals cover the truth and compute the envelope measure, ENV:

$$ENV = \frac{\sum_{i=1}^{k} \mathbf{1}(\hat{N}_{lo}(\omega_i) \leq N(\omega_i) \leq \hat{N}_{up}(\omega_i))}{k} \tag{16}$$

We first simulated 3 datasets of $n = 10$ individuals with an average number of 100 segregating sites under different types of population size trajectories: constant, exponential growth and instantaneous growth. Results are depicted in Figure 5. Posterior medians and 95% credible intervals are shown as black curves and gray shaded areas respectively. The trajectory used to simulate the data is depicted as a dashed line. Figure 5 shows that our BESTT method recovers the constant and exponential growth trajectories very well but the instantaneous growth scenario is less accurate and with high uncertainty (wide credible intervals). In all three cases, our envelope measure is above 95%. Performance measures on all simulations are summarized in table 1.

We analyzed the effect of increasing the number of segregating sites, the number of samples and the number of independent genealogies on posterior inference with BESTT. In all three cases, we

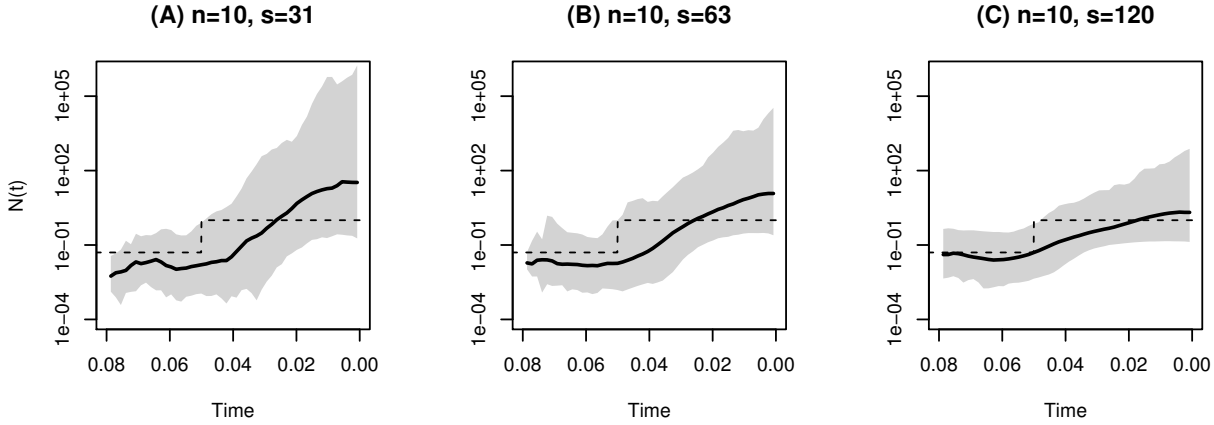


Figure 6: **Varying the number of segregating sites.** Posterior inference from simulated data of $n = 10$ sequences under a population size trajectory with instantaneous growth (dashed lines). s is the number of segregating sites. Posterior medians are depicted as solid black lines and 95% Bayesian credible intervals are depicted by shaded areas.

expect our method to better recover the truth. Figure 6 shows our results on simulated data under a population size trajectory with instantaneous growth (Equation 13) of $n = 10$ individuals with 31, 63 and 120 segregating sites. As expected, our method recovers the truth with higher precision (MRW) and accuracy (SRE) when we increase the number of segregating sites. Increasing the number of segregating sites may result in more constraints in the gene tree. For $n = 10$, there are 7936 possible ranked tree shapes, however for the datasets simulated with 31, 63 and 102 segregating sites, there are only 2582 ± 32, 2670 ± 34 and 556 ± 7 ranked tree shapes compatible with their corresponding gene trees. These numbers were estimated by importance sampling (Cappello and Palacios, 2019).

Table 1: Empirical measures of performance in the simulations described in the text

Simulation	% ENV	SRE	MRW
Instantaneous growth(n=10,s=31)	96	5.87	124352
Instantaneous growth (n=10,s=63)	100	2.15	2296
Instantaneous growth (n=10,s=120)	98	0.53	80
Instantaneous growth (n=25)	90	0.40	3.43
Instantaneous growth (n=35)	92	0.31	3.16
Constant	100	0.30	1.16
Exponential	100	0.35	5.45
Exp. & const. (n=10, 1 locus)	100	4.31	22608
Exp. & const. (n=10, 5 loci)	100	2.37	309.1
Exp. & const. (n=10, 10 loci)	100	0.16	4.19

As another performance assessment, we simulated datasets from a population size trajectory

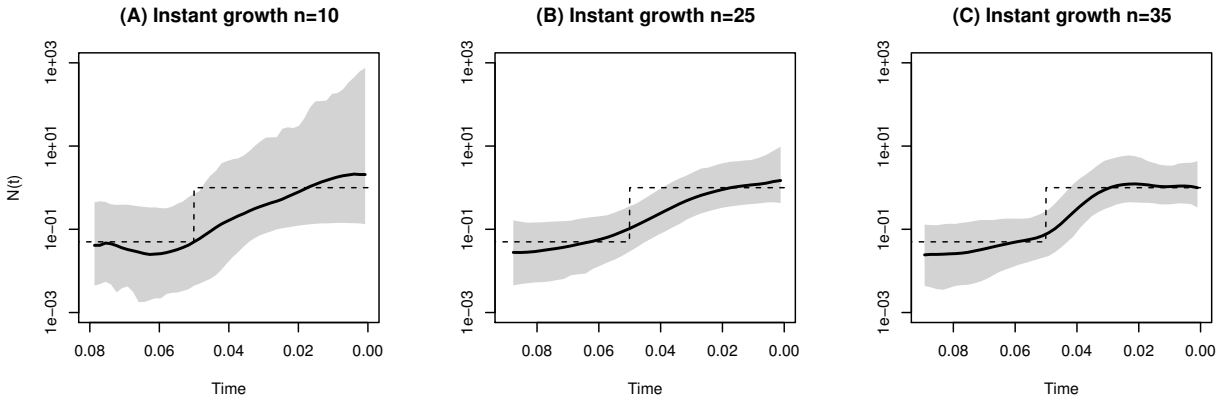


Figure 7: **Varying the number of samples under a population size trajectory with instantaneous growth.** Posterior inference from simulated data of $n = 10$, 25 and 35 sequences under the population size trajectory with instantaneous growth. Shaded areas correspond to 95% credible intervals, solid lines to posterior median and dashed line to the truth.

with instantaneous growth with varying number of samples. We simulated datasets with $n = 10$, 25 and 35 samples with 215 expected number of segregating sites. Our results depicted in Figure 7 show that our method performs better in terms of SRE and MRE when the number of samples increases. Similarly, precision (MRW) and accuracy (SRE) increases when inference is done from a larger number of independent datasets. Finally, Figure 8 shows our results from 1, 5 and 10 datasets simulated from 1, 5 and 10 independent genealogies of 10 individuals with a population size trajectory of growth followed by a constant period (Equation 12). As expected, our method's performance substantially increases by increasing the number of genealogies.

3.2 Comparison to other methods

To our knowledge, there is no other method for inferring (variable) effective population size over time from haplotype data that assumes the infinite sites mutation and a nonparametric prior on $N(t)$, therefore we cannot have a direct comparison of our method to others. Moreover, our method is the only one that explicitly averages over Tajima genealogies instead of Kingman genealogies. BEAST (Drummond et al., 2012) is a program for analyzing molecular sequences that uses MCMC to average over the Kingman tree space and it is therefore a good reference for comparison to our method. We compared our results to the Extended Bayesian Skyline method (Heled and Drummond, 2008) implemented in BEAST.

Since the infinite sites mutation model is not implemented in BEAST, we first converted our simulated sequences of 0s and 1s to sequences of nucleotides by sampling s ancestral nucleotides uniformly on $\{A, T, C, G\}$ and assigning one of the remaining 3 types uniformly at random to be the mutant type. This corresponds to a simulation of the Jukes-Cantor mutation model (Jukes and Cantor, 1969) that is currently implemented in BEAST.

We compare the results of BESTT depicted in Figure 5 to those of BEAST (Drummond et al., 2005, 2012) in Figure 9. We note that results from BEAST are generated from 10 million iterations and thinned every 1000 iterations, while results from BESTT are generated from 200 thousand iterations.

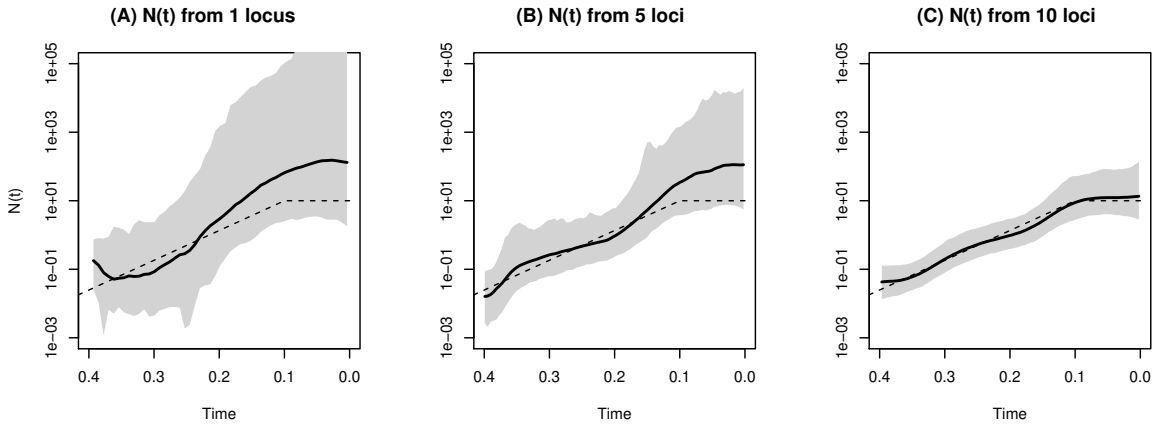


Figure 8: **Multiple independent datasets.** Posterior inference from simulated data of $n = 10$ sequences under exponential followed by constant trajectory (eq. 12). **(A)** Inference from a single simulated dataset, **(B)** from 5 independently simulated datasets, and **(C)** from 10 independently simulated datasets. Shaded areas correspond to 95% credible intervals, solid black lines show posterior medians and dashed lines show the simulated truth.

We compared our point estimates $\hat{N}(t)$ from both methods to the ground truth for each simulation (Table 2). In the three cases BESTT has better envelope than BEAST. For the exponential growth simulation (Figure 9, second row) the BEAST result has better SRE and MRW, however, the credible intervals are uneven with very wide intervals at the ends. For the instantaneous growth simulation (Figure 9, third row), BEAST does not provide results beyond the time point 0.06, for this reason we recomputed the performance statistics for the overlapping time interval $(0, 0.06)$. In this interval, BESTT outperforms BEAST in terms of envelope and SRE.

Table 2: Performance comparison between BESTT and BEAST in simulations

	% ENV		SRE		MRW	
Simulation	BESTT	BEAST	BESTT	BEAST	BESTT	BEAST
Constant	**100**	**100**	**0.3**	0.24	**1.16**	1.49
Exponential	**100**	97	0.35	**0.26**	5.45	**2.56**
Instantaneous growth $(0, 0.06)$	**97**	94	**0.61**	2.65	105.6	**14.95**

Other methods implemented in BEAST are more comparable to BESTT such as Bayesian Skyride (Minin et al., 2008) and Bayesian Skygrid (Gill et al., 2013). These methods assume Gaussian process priors on log $N(t)$ as BESTT, however, for the simulations shown in Figure 9, we were not able to obtain reliable results given that the acceptance probability of the effective population size samplers in BEAST was 0.

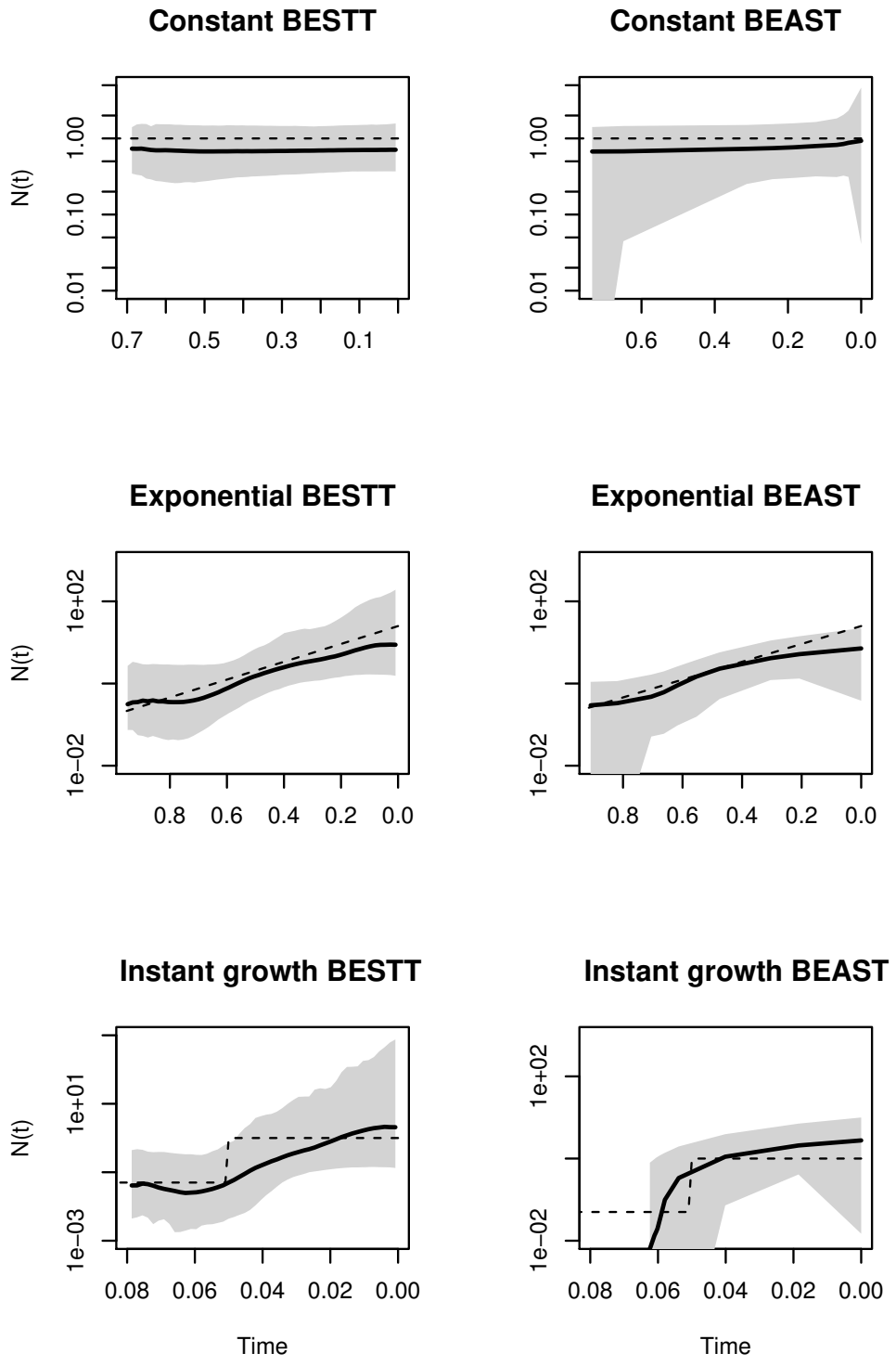


Figure 9: **BESTT and BEAST Comparison.** Posterior inference from simulated data of $n = 10$ sequences under constant, exponential and instantaneous growth trajectories (rows) obtained from our method BESTT (first column) and BEAST (second column). Shaded areas correspond to 95% credible intervals, solid black lines show posterior medians and dashed lines show the simulated truth.

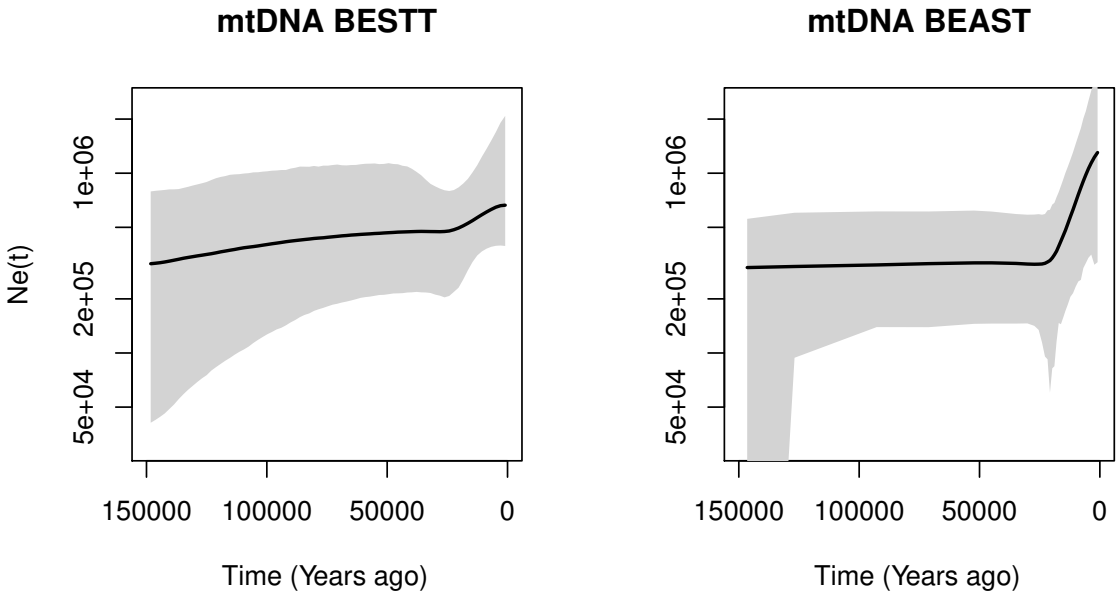


Figure 10: Posterior inference of female effective population size from 35 mtDNA samples from Yoruban individuals in the 1000 Genomes Project using our method BESTT (first plot) and the BEAST Extended Bayesian Skyline Plot (second plot). Posterior median curves are depicted as solid black lines and 95% credible intervals by shaded regions.

4 Inferring human population demography from mtDNA

We selected $n = 35$ samples of mtDNA at random from 107 Yoruban individuals available from the 1000 Genomes Project phase 3 (The 1000 Genomes Project Consortium, 2015). We retained the coding region: base pairs $576 - 16,024$ according to the rCRS reference of Human Mitochondrial DNA (Anderson et al., 1981; Andrews et al., 1999) and removed 38 indels. Of the 260 polymorphic sites, we retained 240 sites compatible with the infinite sites mutation model. The final file is available in `https://github.com/JuliaPalacios/phylodyn`. To encode our data as 0s and 1s, we use the inferred root sequence **RSRS** of Behar et al. (2012) to define the ancestral type at each site. To rescale our results in units of years, we assumed a mutation rate per site per year of 1.3×10^{-8} (Rebolledo-Jaramillo et al., 2014). We compare our results with the Extended Bayesian Skyline method (Drummond et al., 2012) implemented in BEAST in Figure 10. When applying BEAST, we assumed the Jukes-Cantor mutation model. Both methods detect an inflection point around 20kya followed by exponential growth. The mean time to the most recent ancestor (TMRCA) inferred for these YRI mtDNA samples with BESTT is around 170kya with a 95% BCI of $(142868, 207455)$, while the mean TMRCA inferred with BEAST is around 160kya with a 95% BCI of $(133239, 196900)$. In Appendix B, we include two more comparisons of BESTT and BEAST.

5 Discussion

The size of emergent sequencing datasets prohibits the use of standard coalescent modeling for inferring evolutionary parameters. The main computational bottleneck of coalescent-based inference of evolutionary histories lies in the large cardinality of the hidden state space of genealogies. In the standard Kingman coalescent, a genealogy is a random labeled bifurcating tree that models the set of ancestral relationships of the samples. The genealogy accounts for the correlated structure induced by the shared past history of organisms and explicit modeling of genealogies is fundamental for learning about the past history of organisms. However, the genomic era is producing large datasets that require more efficient approaches that efficiently integrate over the hidden state space of genealogies.

In this manuscript we show that a lower resolution coalescent model on genealogies, the "Tajima's coalescent", can be used as an alternative to the standard Kingman coalescent model. In particular, we show that the Tajima coalescent model provides a feasible alternative that integrates over a smaller state space than the standard Kingman model. The main advantage in Tajima's coalescent is to model the ranked tree topology as opposed to the fully labeled tree topology as in Kingman's coalescent.

A priori, the cardinality of the state space of ranked tree shapes is much smaller than the cardinality of the state space of labeled trees. However, in this manuscript we show that when the Tajima coalescent model is coupled with the infinite sites mutation model, the space of ranked tree shapes is constrained by the data and the reduction on the cardinality of the hidden state space of Tajima's trees is even more pronounced than expected.

In order to leverage the constraints imposed by the data and the infinite-sites mutation model, we apply Dan Gusfield's perfect phylogeny algorithm (Gusfield, 1991) to represent sequence alignments as a gene tree. We exploit the gene tree representation for conditional likelihood calculations and for exploring the state space of ranked tree shapes.

For the calculation of the likelihood of the data conditioned on a given Tajima's genealogy, we augment the gene tree representation of the data with the Tajima's genealogy and map observed mutations to branches. We define a directed acyclic graph (DAG) with the augmented gene tree. This new representation as a DAG allows for calculating the likelihood as a depth-first search algorithm that transverses the gene tree from the leaves to the root. Our present implementation computational's bottleneck lies in the likelihood calculation. In future studies, our proposed algorithm for likelihood calculation can be further optimized as a sum-product algorithm; however, we are able to infer effective population size trajectories from samples of size $n \approx 35$ in a regular personal laptop computer within few hours.

Our statistical framework draws on Bayesian nonparametrics. We place a flexible geometric random walk process prior on the effective population size that allows us to recover population size trajectories with abrupt changes in simulations. The inference procedure proposed in this manuscript relies on Markov chain Monte Carlo (MCMC) methods with 3 large Gibbs block updates of: coalescent times, effective population size trajectory and ranked tree shape topology. We use Hamiltonian Monte Carlo updates for continuous random variables: coalescent times and effective population size; and a Metropolis Hastings sampler for exploring the space of ranked tree shapes. For exploring the genealogical space, Markovtsova et al. (2000) suggest a joint local proposal for both coalescent times and topology. Here we restrict our attention to the topology alone. A future line of research includes the development of a joint local proposal of coalescent times and ranked tree shapes. We also envision that a joint sampler of coalescent times and effective population size

trajectories should improve mixing and convergence.

Finally, haplotype data of many organisms is usually sparse with few unique haplotypes presented at high frequencies. Our proposed method is ideally suited for this scenario where the space of ranked tree shapes is drastically smaller than the space of labeled topolgies.

Acknowledgements

This research is supported in part by a National Institutes of Health grant R01- GM-131404 and the Alfred P. Sloan Foundation to J.A.P.. We want to acknowledge the developers of `R-ape`, `R-phangorn` and `R-phylodyn` that facilitated our implementations. A.V. was supported in part by the chaire program Modélisation Mathématique et Biodiversité of Veolia Environnement - École Polytechnique - Museum National d'Histoire Naturelle - Fondation X. A.V. and J.A.P. was supported by the France-Stanford Center for interdisciplinary Studies. This was work also supported by the National Science Foundation CAREER Award DBI-1452622 to S.R.

References

Aberer, A. J., Stamatakis, A., and Ronquist, F. (2016). An efficient independence sampler for updating branches in bayesian markov chain monte carlo sampling of phylogenetic trees. Systematic Biology, 65(1):161.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., et al. (1981). Sequence and organization of the human mitochondrial genome. Nature, 290(5806):457.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial dna. Nature genetics, 23(2):147.

Behar, D. M., Van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N. M., Kivisild, T., Torroni, A., and Villems, R. (2012). A "copernican" reassessment of the human mitochondrial dna tree from its root. The American Journal of Human Genetics, 90(4):675–684.

Cappello, L. and Palacios, J. A. (2019). Sequential importance sampling for multi-resolution kingman-tajima coalescent counting. arXiv preprint arXiv:1902.05527.

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

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. Annual Review of Genetics, 29(1):401–421.

Drummond, A. and Rodrigo, A. (2000). Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Molecular Biology and Evolution, 17(12):1807–1815.

Drummond, A., Suchard, M., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29:1969–1973.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution, 22(5):1185–1192.

Gill, M., Lemey, P., Faria, N., Rambaut, A., Shapiro, B., and Suchard, M. (2013). Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. Molecular Biology and Evolution, 30:713–724.

Griffiths, R. and Tavaré, S. (1994a). Simulating probability distributions in the coalescent. Theoretical Population Biology, 46(2):131 – 159.

Griffiths, R. and Tavaré, S. (1995). Unrooted genealogical tree probabilities in the infinitely-many-sites model. Mathematical Biosciences, 127(1):77 – 98.

Griffiths, R. C. and Tavaré, S. (1994b). Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 344(1310):403–410.

Griffiths, R. C. and Tavaré, S. (1996). Monte Carlo inference methods in population genetics. Mathematical and Computer Modelling, 23(8-9):141–158.

Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees. Networks, 21(1):19–28.

Heled, J. and Drummond, A. (2008). Bayesian inference of population size history from multiple loci. BMC Evolutionary Biology, 8(1):1–289.

Hobolth, A., Uyenoyama, M. K., and Wiuf, C. (2008). Importance sampling for the infinite sites model. Statistical applications in genetics and molecular biology, 7.

Jukes, T. H. and Cantor, R. C. (1969). Evolution of protein molecules. In Mammalian Protein Metabolism, pages 21–132. Academic, New York.

Karcher, M. D., Palacios, J. A., Lan, S., and Minin, V. N. (2017). phylodyn: an r package for phylodynamic simulation and inference. Molecular Ecology Resources, 17(1):96–100.

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

Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations. In Koch, G. and Spizzichino, F., editors, Exchangeability in Probability and Statistics, pages 97–112. North-Holland, Amsterdam.

Kuhner, M. and Smith, L. (2007). Comparing likelihood and Bayesian coalescent estimation of population parameters. Genetics, 175(1):155–165.

Kuhner, M. K. (2006). LAMARC 2.0: maximum likelihood and bayesian estimation of population parameters. Bioinformatics, 22(6):768.

Kuhner, M. K., Yamato, J., and Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. Genetics, 149(1):429–434.

Lan, S., Palacios, J. A., Karcher, M., Minin, V., and Shahbaba, B. (2014). An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. CoRR, abs/1126456.

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

Markovtsova, L., Marjoram, P., and Tavaré, S. (2000). The age of a unique event polymorphism. Genetics, 156(1):401–409.

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

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

Palacios, J. A., Wakeley, J., and Ramachandran, S. (2015). Bayesian nonparametric inference of population size changes from sequential genealogies. Genetics.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. Genetics, 164(4):1645–1656.

Rebolledo-Jaramillo, B., Su, M. S.-W., Stoler, N., McElhoe, J. A., Dickins, B., Blankenberg, D., Korneliussen, T. S., Chiaromonte, F., Nielsen, R., Holland, M. M., Paul, I. M., Nekrutenko, A., and Makova, K. D. (2014). Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial dna. Proceedings of the National Academy of Sciences, 111(43):15474–15479.

Sainudiin, R., Stadler, T., and Véber, A. (2015). Finding the best resolution for the kingman-tajima coalescent: theory and applications. Journal of Mathematical Biology, pages 1–41.

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

Schliep, K. (2011). phangorn: phylogenetic analysis in R. Bioinformatics, 27(4):592–593.

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

Slatkin, M. and Hudson, R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics, 129(2):555–562.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(4):605–635.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature, 526:68–74.

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

Whidden, C. and Matsen, IV, F. A. (2015). Quantifying mcmc exploration of phylogenetic tree space. Systematic Biology, 64(3):472.

Wu, Y. (2010). Exact computation of coalescent likelihood for panmictic and subdivided populations under the infinite sites model. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(4):611–618.

6 Appendix A

Markovian proposal of ranked tree shapes. The following algorithm generates a new ranked tree shape from a Markovian proposal and outputs the corresponding transition probabilities. This proposal is used in section 2.8.1.

Algorithm 1 Transition proposals for ranked tree shapes

Input: $\mathbf{F}_n$
Output: $\mathbf{F}_n^*$, $q(\mathbf{F}_n \mid \mathbf{F}_n^*)$, $q(\mathbf{F}_n^* \mid \mathbf{F}_n)$

1. Set $\mathbf{F}_n = \mathbf{F}_n^*$.
2. Sample with uniform discrete probability a coalescent event k from the set $\{1, \ldots, n-2\}$. Set $q_1 = \frac{1}{n-2}$.
3. **If** vintage k coalesces in the coalescent event $k+1$

   If the lineages coalescing in k are leaves

   (a) Update $\mathbf{F}_n^*$: merge one of the lineages of vintage k with the lineage coalescing at $k+1$ at time k, then merge the second lineage at time $k+1$, set $q_2 = 1$

   (b) Compute the probability q_2' of restoring the ordering of $\mathbf{F}_n^*$ to $\mathbf{F}_n$.

   Else

   (a) Sample one the lineages of vintage k with uniform discrete probability. Set $q_2 = \frac{1}{2}$

   (b) update $\mathbf{F}_n^*$: merge the sampled lineage with the one coalescing at time $k+1$. At time $k+1$, merge the lineage not sampled with the new vintage k.

   (c) Compute the probability q_2' of restoring the ordering of $\mathbf{F}_n^*$ to $\mathbf{F}_n$.

   Else

   Swap the labels of vintages k and $k+1$. Set $q_2 = 1$ and $q_2' = 1$.
4. $q(\mathbf{F}_n^* \mid \mathbf{F}_n) = q_1\, q_2$, $q(\mathbf{F}_n \mid \mathbf{F}_n^*) = q_1\, q_2'$.

Algorithms for conditional likelihood calculation. The following two algorithms detail the calculation of $\Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu)$. $\mathbf{Y}$ is encoded in $GeneTree$, the observed data as a Tree structure. Each node in $GeneTree$ has number of descendants (or lineages) and mutation information attached to it. Tajima's genealogy $\mathbf{g}^T$ is encoded as $Fpath$ that contains the ranked tree shape $\mathbf{F}_n$ and $times$ that contains the vector of coalescent times $\mathbf{t}$ multiplied by the mutation rate μ.

Algorithm 2 `Calculate Likelihood`($Fpath, times, GeneTree$) procedure

Require: $GeneTree$, $FPath$
Ensure: Log Likelihood LL
1: Initiate *pool* to be the set of leave nodes of $GeneTree$ with at least one descendant. Initiate LL and *index* to be zero. Initiate *current_path* to be empty.
2: Call `CalcLL_recursive`(LL, *index*, *current_path*, $Fpath$, *times*, *Genetree*).
3: **return** LL

Algorithm 3 `CalcLL_recursive`(LL, *index*, *current_path*, $Fpath$, *times*, *Genetree*) procedure

1: **if** $index = len(path)$ {When a complete path node is found} **then**
2: **for** *node* in *tree* **do**
3: Calculate log likelihood based on *times* and number of mutations of *node* in *current_path*.
4: Accumulate to total log likelihood LL
5: **end for**
6: **else**
7: **for** *node* in *pool* **do**
8: Check compatibility of the *node*, according to the given $Fpath$.
9: **if** *node* is compatible with $Fpath$ **then**
10: Update *node* by coalescing two lineages according to current step in $Fpath$
11: Update *pool*. If a *node* does not have more than one lineage to coalesce, remove *node* from pool. If *node* is being removed from pool, update its parent *node*, and potentially add parent node to *pool* if parent node has more than 2 lineages.
12: Append this *node* to *current_path*
13: Call `CalcLL_recursive`(LL, $index + 1$, *current_path*, $Fpath$, *Genetree*)
14: Restore previous *node*, *pool* and *current_path*
15: **end if**
16: **end for**
17: **end if**

7 Appendix B

We replicated the BEAST EBSP Analysis of the 35 Yoruban individuals from the 1000 Genomes Project phase 3 using the whole mtDNA coding region consisting of 15409 sites. In both cases we assumed the Jukes-Cantor mutation model (Jukes and Cantor, 1969). Figure 11 shows the comparison between EBSP inference from the 240 segregating sites retained in section 4 that are compatible with the infinite sites mutation model assumption. In both cases we recover very similar trajectories.

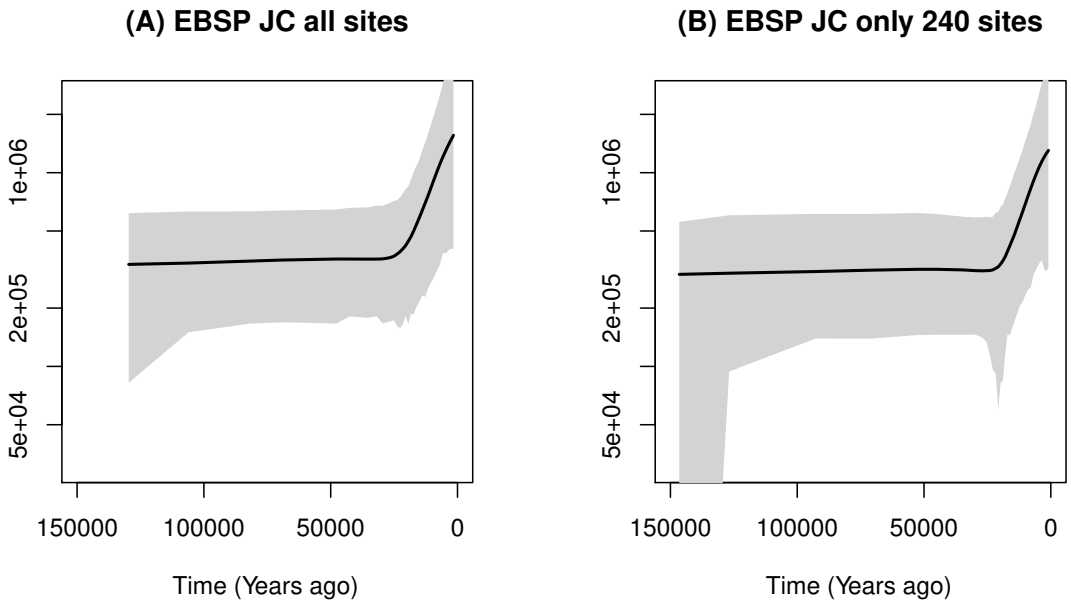


Figure 11: Posterior inference of female effective population size from 35 mtDNA samples from Yoruban individuals in the 1000 Genomes Project using BEAST EBSP (first plot) from all 15409 sites and the BEAST EBSP (second plot) from the 240 segregating sites retained. In both cases, the mutation model assumed is Jukes Cantor (JC). Posterior median curves are depicted as solid black lines and 95% credible intervals by shaded regions.

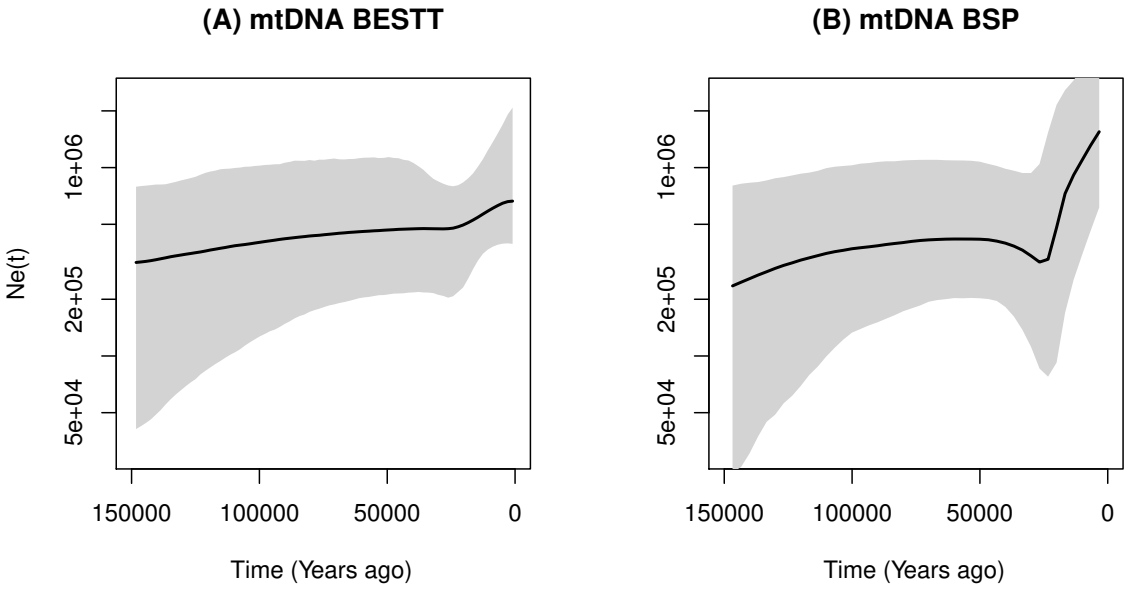


Figure 12: Posterior inference of female effective population size from 35 mtDNA samples from Yoruban individuals in the 1000 Genomes Project using our BESTT (first plot) from only 240 segregating sites and the BEAST BSP (second plot) from all the 15409 sites. Posterior median curves are depicted as solid black lines and 95% credible intervals by shaded regions.

In addition, we compared our results with BEAST Bayesian Skyline Plot (BSP) (Drummond and Rodrigo, 2000). For our reduced dataset of 240 segregating sites, we could not generate valid inference of $N(t)$ with Metropolis-Hastings acceptance probability greater than 0. Instead we were able to generate results with BEAST BSP from the complete dataset of 15409 sites. The comparison of our method from 240 segregating sites to BEAST BSP from 15409 sites is depicted in Figure 12.