

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

Julia A. Palacios<sup>1,2,\*</sup>, Amandine Véber<sup>3</sup>, Lorenzo Cappello<sup>1</sup>, Zhangyuan Wang<sup>4</sup>, John Wakeley<sup>5</sup>  
and Sohini Ramachandran<sup>6,7</sup>

\*corresponding author: [juliapr@stanford.edu](mailto: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,

<sup>1</sup>Department of Statistics. Stanford University, Stanford, CA 94305.

<sup>2</sup>Department of Biomedical Data Science. Stanford School of Medicine, Stanford, CA 94305.

<sup>3</sup>CMAP, École Polytechnique, CNRS, Palaiseau, France

<sup>4</sup>Department of Computer Science. Stanford University, Stanford, CA 94305

<sup>5</sup>Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

<sup>6</sup>Center for Computational Molecular Biology, Brown University, Providence, RI 02912

<sup>7</sup>Department of Ecology and Evolutionary Biology, Brown University, Providence, RI 02912

28 1991; Donnelly and Tavaré, 1995). Our goal is to infer  $(N(t))_{t \geq 0}$  which we will refer to as the  
29 “effective population size trajectory”.

30 Multiple methods have been developed to infer  $(N(t))_{t \geq 0}$  using the standard coalescent model  
31 on genomic sequence datasets (Griffiths and Tavaré, 1996; Kuhner and Smith, 2007; Minin et al.,  
32 2008; Li and Durbin, 2011; Drummond et al., 2012; Palacios and Minin, 2013; Gill et al., 2013;  
33 Sheehan et al., 2013; Palacios et al., 2015). These methods must contend with two major challenges:  
34 (i) choosing a prior distribution or functional form for  $(N(t))_{t \geq 0}$ , and (ii) integrating over the  
35 large hidden state space of genealogies. For example, several previous approaches have assumed  
36 exponential growth (Griffiths and Tavaré, 1996; Kuhner et al., 1998; Kuhner and Smith, 2007), in  
37 which case the estimation of  $(N(t))_{t \geq 0}$  is reduced to the estimation of one or two parameters. In  
38 general, the functional form of  $(N(t))_{t \geq 0}$  is unknown and needs to be inferred. A commonly used  
39 naive nonparametric prior on  $(N(t))_{t \geq 0}$  is a piecewise linear or constant function defined on time  
40 intervals of constant or varying sizes (Heled and Drummond, 2008; Sheehan et al., 2013; Schiffels  
41 and Durbin, 2014). The specification of change points in such time-discretized effective population  
42 size trajectories is inherently difficult because it can lead to runaway behavior or large uncertainty  
43 in  $(\hat{N}(t))_{t \geq 0}$  (Minin et al., 2008; Heled and Drummond, 2008; Li and Durbin, 2011; Sheehan et al.,  
44 2013; Palacios et al., 2015). These difficulties can be avoided by the use of Gaussian-process priors  
45 in a Bayesian nonparametric framework, allowing accurate and precise estimation (Palacios and  
46 Minin, 2013; Gill et al., 2013; Lan et al., 2014; Palacios et al., 2015).

47 The second challenge for coalescent-based inference of  $(N(t))_{t \geq 0}$  is the integration over the hid-  
48 den state space of genealogies. Given molecular sequence data  $\mathbf{Y}$  and a mutation model with  
49 vector of parameters  $\boldsymbol{\mu}$ , current methods rely on calculating the marginal likelihood function  
50  $\Pr(\mathbf{Y} | (N(t))_{t \geq 0}, \boldsymbol{\mu})$  by integrating over all possible coalescence and mutation events. Under the  
51 infinite-sites mutation model without intra-locus recombination (Watterson, 1975), this integra-  
52 tion requires a computationally expensive importance sampling technique or Markov Chain Monte  
53 Carlo (MCMC) techniques (Griffiths and Tavaré, 1994a; Stephens and Donnelly, 2000; Hobolth  
54 et al., 2008; Wu, 2010). Moreover, a maximum likelihood estimate of  $(N(t))_{t \geq 0}$  cannot be ex-  
55 plicitly obtained; instead, it is obtained by exploring a grid of parameter values. For finite-sites  
56 mutation models, current methods approximate the marginal likelihood function by integrating  
57 over all possible genealogies via MCMC methods (Equation (1); Kuhner (2006); Drummond et al.  
58 (2012)). Both cases may be represented as

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

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

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

71 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  
72 the Tajima’s genealogy of the sample (see Figure 1A and Section 2.4),  $(\log N(t))_{t \geq 0}$  has Gaussian  
73 process prior with precision hyperparameter  $\tau$ , and mutations occur according to the infinite-sites  
74 model of Watterson (1975). This results in a new more efficient method for inferring  $(N(t))_{t \geq 0}$   
75 called **B**ayesian **E**stimation by **S**ampling **T**ajima’s **T**rees (BESTT), with a drastic reduction in  
76 the state space of genealogies. We show using simulated data that BESTT can accurately infer  
77 effective population size trajectories and that it provides a more efficient alternative than Kingman’s  
78 coalescent models.

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

## 88 2 Methods/Theory

### 89 2.1 Overview of BESTT

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

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

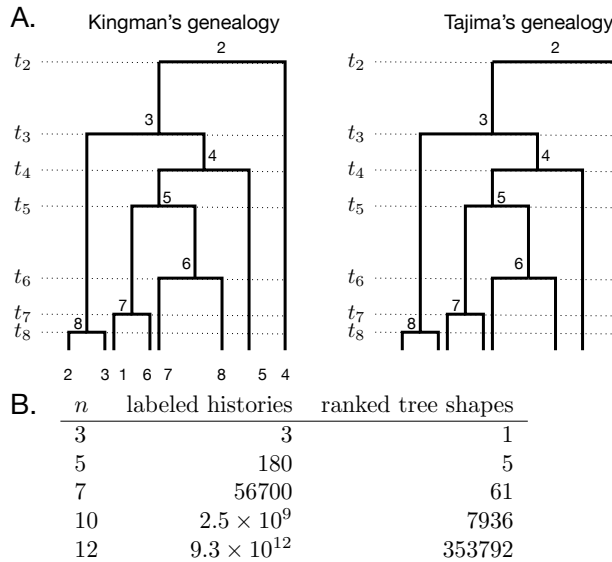


Figure 1: **For a sample of size  $n$ , the number of Tajima's genealogies is superexponentially fewer compared to the number of Kingman's genealogies.**

**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,  $n$ .

## 113 2.2 Summarizing sequence data $\mathbf{Y}$ as haplotypes and mutation groups

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

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

## 130 2.3 Representing $\mathbf{Y}_{h \times m}$ as a gene tree

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

137 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  
138  $h$  leaves such that (Figure 2B):

- 139 1. Each row of  $\mathbf{Y}_{h \times m}$  corresponds to exactly one leaf of  $\mathcal{T}$ . The black numbers at leaf nodes in  
140 Figure 2B are the haplotype frequencies.
- 141 2. Each mutation group  $m$  of  $\mathbf{Y}_{h \times m}$  is represented by exactly one edge of  $\mathcal{T}$ . The red numbers  
142 along edges in Figure 2B give the cardinality of each mutation group (*i.e.* the number of  
143 segregating sites in each mutation group; see Figure 2A). The edges of  $\mathcal{T}$  corresponding  
144 to each mutation group are labeled accordingly (letters in Figures 2A and 2B). Note that  
145 external edges need not be labeled if there are no mutations exclusively present in the group  
146 of individuals descending from those edges.
- 147 3. The labels and the numbers associated with the edges along the unique path from the root  
148 to a leaf exactly specify a row of  $\mathbf{Y}_{h \times m}$ .

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

## 151 2.4 Tajima’s genealogies

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

163 The process that keeps track of the blocks formed at each time step (including the labels of  
164 the individuals of the sample within each block, which share a common ancestor) induces a ranked  
165 labeled rooted binary tree that we call a “labeled topology”. Note that when the labeled tree topol-  
166 ogy is presented together with the coalescent times, ranking of coalescent events is redundant. For  
167 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  $\mathbf{t} = (t_n, \dots, t_2)$  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,  $t_n, \dots, t_2$ , 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  $\Pr(\mathbf{g} | (N(t))_{t \geq 0})$ , can be factored as the  
product of the probability of Kingman’s genealogy and the coalescent times density:

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

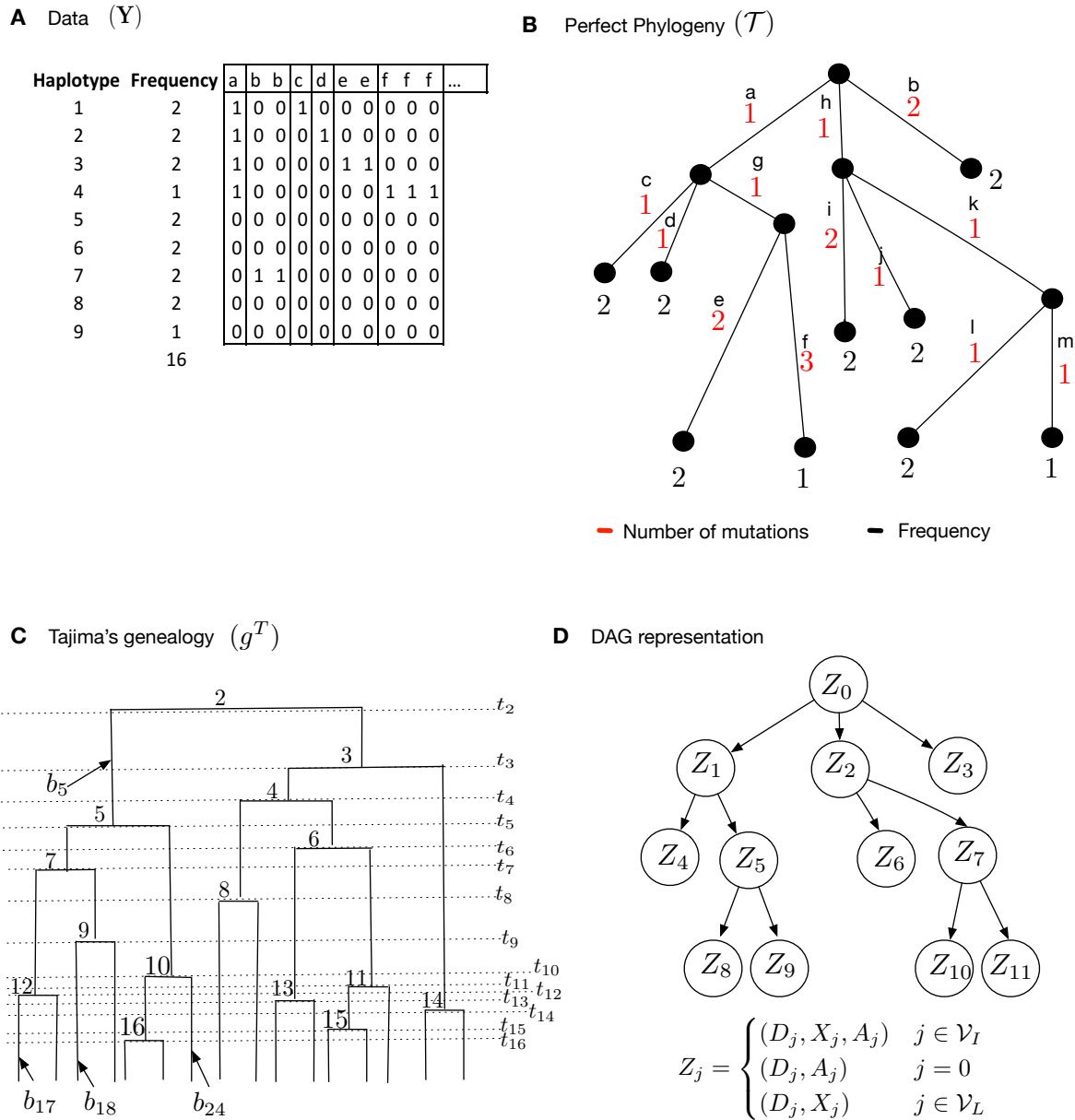


Figure 2: **Data structures exploited by our method, BESTT, for calculating the conditional likelihood of the data.** **A.** Compressed data representation  $Y_{h \times m}$  of  $n = 16$  sequences and  $s = 18$  (columns, only the first 10 of which are shown), comprised of 9 haplotypes and 13 mutation groups. Rows correspond to haplotypes and each polymorphic site is labeled by its mutation group  $\{a, b, c, \dots, m\}$ . **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.

168 where  $\Pr(K_n) = 2^{n-1}/[n!(n-1)!]$ . We refer to the distribution of a genealogy as density, although  
 169 the distribution is a mixed distribution of continuous random variables (coalescent times) and dis-  
 170 crete random variables (topology). Here again, we use the notation  $\Pr(\cdot)$  to denote the probability  
 171 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 ( $i = 2, \dots, n$  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  $\{2, 3, \dots, n\}$ . 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  $\{2, 3, \dots, n\}$ ,  $\mathbf{F}_{n,i}$  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  $\Pr(\mathbf{g}^T | (N(t))_{t \geq 0})$ , can be factored as the product of the probability of Tajima's genealogy and the coalescent times density:

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

172 with

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

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

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

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

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

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

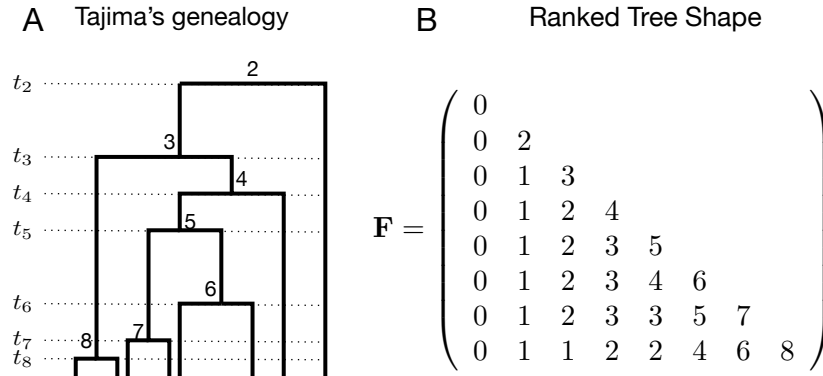


Figure 3: **Ranked tree shape** Left: Example of a Tajima's genealogy (redrawn from Figure 1A) with coalescent events ranked from 2 at time  $t_2$  to  $n$  at time  $t_n$ . Right: The corresponding  $\mathbf{F}_n$  matrix, with  $n = 8$ , that encodes the ranked tree shape information of the Tajima's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

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

## 190 2.5 An augmented data representation using directed acyclic graphs

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

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

203 **Constructing the DAG  $\mathbf{D}$ .** Our DAG  $\mathbf{D} = \{\mathbf{Z}, E\}$  (Figure 2D) with nodes  $\mathbf{Z}$  and edges  $E$   
 204 is constructed from a gene tree  $\mathcal{T}$ . The number of internal nodes in the DAG  $\mathbf{D}$  is the same as  
 205 the number of internal nodes in  $\mathcal{T}$ . However, sister leaf nodes in  $\mathcal{T}$  with the same number of  
 206 descendants are grouped together in  $\mathbf{D}$ . For example, the leaves in Figure 2B subtending from  
 207 edges  $i$  and  $j$  are grouped into  $Z_6$  in Figure 2D, as they both have haplotype frequency 2. However,  
 208 the leaves subtending from the  $e$  and  $f$  edges are not grouped (and correspond to  $Z_8$  and  $Z_9$  in the  
 209 DAG Figure 2D) since they have respective haplotype frequencies 2 and 1. We label the root node  
 210 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  
 211  $i < j$ , we assign a directed edge  $E_{i,j}$  if the node in  $\mathcal{T}$  corresponding to  $Z_i$  is connected to the node



212 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  
 213 of leaf nodes is denoted by  $\mathcal{V}_L$ .

**Random variables in  $\mathbf{D}$ .** 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:

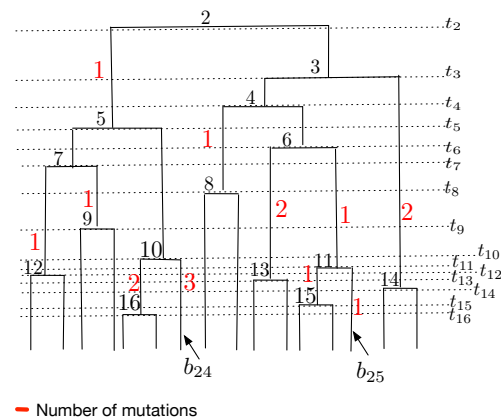
$$Z_j = \begin{cases} (D_j, X_j, A_j) & j \in \mathcal{V}_I, \\ (D_j, A_j) & j = 0 \text{ (the root node)}, \\ (D_j, X_j) & j \in \mathcal{V}_L, \end{cases}$$

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

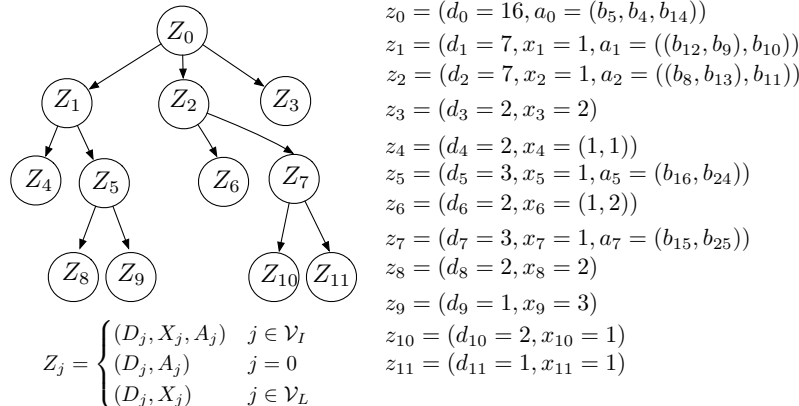
223 **Allocation of mutation groups along  $\mathbf{g}^T$ .** The allocation random variables  $\{A_j\}$  are con-  
 224 strained by the information in the Tajima's genealogy  $\mathbf{g}^T$ . In a given  $\mathbf{g}^T$ , every subtree is labeled  
 225 by its ranking from past to present (Figure 3).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

234 Allocations of mutations to branches are usually not unique and computation of the conditional  
 235 likelihood  $\Pr(\mathbf{Y} \mid \mathbf{g}^T, \mu)$  requires summing over all possible allocations. In Figure 4A we show one  
 236 such possible allocation of the mutation groups of the gene tree in Figure 2B along the Tajima's  
 237 genealogy in Figure 2C. For example, mutation group "a" in Figure 2B with cardinality 1 (number  
 238 in red) is a mutation observed in 7 individuals (sum of black numbers of leaves descending from  
 239 edge marked a). This same mutation group, "a", is shown as a red number 1 in Figure 4A allocated  
 240 to branch  $b_5$ . If  $Z_j$  is an internal node, the number of mutations  $X_j$  separating it from its parent  
 241 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  
 242 length of  $X_j$  is the number of the corresponding sister nodes in  $\mathcal{T}$  that were grouped together in  
 243 forming node  $Z_j$ .  $A_j = (A_{j,1}, \dots, A_{j,|ch(j)|})$  denotes a collection of  $|ch(j)|$  vectors of branch labels  
 244 in  $\mathbf{g}^T$  subtending the child-node subtrees of node  $Z_j$ .  $A_{j,1}$  corresponds to the branch subtending

**A** Tajima's genealogy and a possible allocation of the mutations observed in the data



**B** DAG corresponding to **A**



**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  $X_j = (X_{j,1}, X_{j,2}, \dots, X_{j,s_j})$  that denotes the number of mutations along the branches that subtends from the tree node  $j$  that have  $D_j$  descendants.

245 from the leftmost child node of  $Z_j$  on **D**,  $A_{j,2}$  corresponds to the branch subtending from the next  
 246 child node of  $Z_j$ , etc., and  $A_{j,|ch(j)|}$  corresponds to the branch subtending from the rightmost child  
 247 node of  $Z_j$  on **D**. Observe that, since we group some of the leaf nodes in  $\mathcal{T}$  into a single node in  
 248 **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  
 249 4B.

## 250 **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  $\mu$ . 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 **D**; that is, we

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

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

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  $i \in \{1, \dots, n_I + n_L\}$ , we have

$$\begin{aligned} &\Pr(Z_i = z_i \mid z_{pa(i)}, \mathbf{g}^T, \mu) \\ &= \begin{cases} \Pr(X_i = x_i \mid a_{pa(i)} = b_j, \mathbf{g}^T, \mu) \propto (\mu l_j)^{x_i} & \text{if } |x_i| = 1, \\ \Pr(X_i = (x_{i1}, \dots, x_{ik}) \mid a_{pa(i)} = (b_{j1}, \dots, b_{jk}), \mathbf{g}^T, \mu) \propto \sum_{s \in \Pi(x_i, k)} \prod_{m=1}^k (\mu l_{j_m})^{s_m} & \text{if } |x_i| = k > 1, \end{cases} \end{aligned}$$

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

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

254 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  
 255  $\{l_3, l_4, l_5, l_6, l_7, l_8\}$ . In this case,  $k_1 = 3$  because there will be 3 branches with 2 mutations,  $k_2 = 1$   
 256 because there will be 1 branch with 0 mutations and  $k_3 = 2$  because there will be 2 branches with  
 257 3 mutations. The number of permutations of  $k = 6$  mutations groups divided into  $m_i = 3$  groups  
 258 with cardinalities 2, 0, 3 of sizes 3, 1, 2 is  $6!/(3!1!2!) = 60$ .

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

## 262 2.7 The case of unknown ancestral states

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

268 When the ancestral types are not known, the data (now denoted  $\mathbf{Y}^0$ ) may be represented as  
 269 an unlabeled *unrooted* gene tree. By the remark following Equation (1) in Griffiths and Tavaré  
 270 (1995), if  $s$  is the number of segregating sites, then there are at most  $s + 1$  unlabeled rooted gene

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

$$\Pr(\mathbf{Y}^0 | \mathbf{g}^T, \mu) = \sum_{i=1}^{\mathcal{R}(Y^0)} P(\mathbf{Y}^{(i)} | \mathbf{g}^T, \mu), \quad (6)$$

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

## 278 2.8 Bayesian inference of the effective population size trajectory

279 Our posterior distribution of interest is

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

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 implementa-  
 tion, we assume a regular geometric random walk prior, that is,  $\gamma_1 = \log N(t_1^*)$ ,  $\dots$ ,  $\gamma_B = \log N(t_B^*)$   
 at  $B$  regularly spaced time points in  $[0, T]$  with

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

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

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

285 To summarize the effective population size trajectory, we compute the posterior median and  
 286 95% credible intervals pointwise at each grid point in  $[0, \hat{T}]$ , where  $\hat{T}$  is the maximum time to the  
 287 most recent common ancestor sampled.

### 288 2.8.1 Metropolis-Hastings updates for ranked tree shapes

289 There is a large literature on local transition proposal distributions for Kingman’s topologies (Kuh-  
 290 ner et al., 1998; Rannala and Yang, 2003; Drummond et al., 2012; Whidden and Matsen, 2015;  
 291 Aberer et al., 2016). In this paper, we adapted the local transition proposal of Markovtsova et al.  
 292 (2000) to Tajima’s topologies. We briefly describe the scheme below and provide a pseudocode  
 293 algorithm in the Appendix (Algorithm 1).

294 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  
 295 shape  $\mathbf{F}^*$  in two steps: (1) we first sample a coalescent interval  $k$  uniformly on  $\{1, \dots, n - 2\}$ .  
 296 Given  $k$ , we focus solely on the coalescent event sampled and the one that follows (thus, the last  
 297 coalescent event cannot be sampled). For step (2), there are two possible scenarios: either vintage  
 298  $k$  undergoes a coalescent at event at  $k + 1$ , or it does not. In the former scenario, we choose a new  
 299 pair of lineages at random to coalesce at  $k$  from the 3 lineages subtending  $k$  and  $k + 1$  (excluding

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

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

### 304 2.8.2 Split Hamiltonian Monte Carlo updates of $(\gamma, \tau)$

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

### 308 2.8.3 Hamiltonian Monte Carlo updates of coalescent times

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

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

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

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

### 326 2.8.4 Local updates of coalescent times

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

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

### 331 2.8.5 Multiple Independent loci

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

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

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

## 339 3 Results

### 340 3.1 The performance of BESTT in applications to simulated data

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

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

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

- 346 2. A trajectory with instantaneous growth:

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

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

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

349 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  
 350 while there are  $j$  lineages, we drew the total number of mutations (segregating sites)  $s$  according  
 351 to a Poisson distribution with parameter  $\mu L$ . We then placed the mutations uniformly at random  
 352 along the branches of the genealogy. For each of the  $s$  mutations, we assigned the mutant type  
 353 to individuals descending from the branch where the mutation occurred and the ancestral type  
 354 otherwise.

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

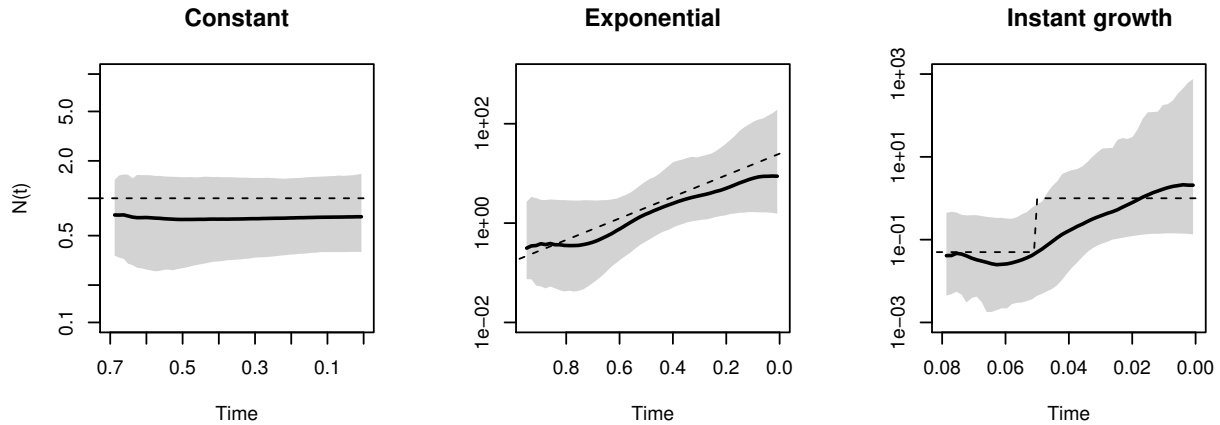


Figure 5: **Different trajectories.** Posterior inference from simulated data of  $n = 10$  sequences under constant, exponential and instantaneous growth population size trajectories (dashed lines). Posterior medians are depicted as solid black lines and 95% Bayesian credible intervals are depicted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

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

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

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

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

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

364 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  
 365 limit of the estimated posterior distribution of  $N(\omega_i)$ . In addition, we measured how well the 95%  
 366 credible intervals cover the truth and compute the envelope measure,  $ENV$ :

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

367 We first simulated 3 datasets of  $n = 10$  individuals with an average number of 100 segregating  
 368 sites under different types of population size trajectories: constant, exponential growth and instan-  
 369 tantaneous growth. Results are depicted in Figure 5. Posterior medians and 95% credible intervals are  
 370 shown as black curves and gray shaded areas respectively. The trajectory used to simulate the data  
 371 is depicted as a dashed line. Figure 5 shows that our BESTT method recovers the constant and  
 372 exponential growth trajectories very well but the instantaneous growth scenario is less accurate and  
 373 with high uncertainty (wide credible intervals). In all three cases, our envelope measure is above  
 374 95%. Performance measures on all simulations are summarized in table 1.

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

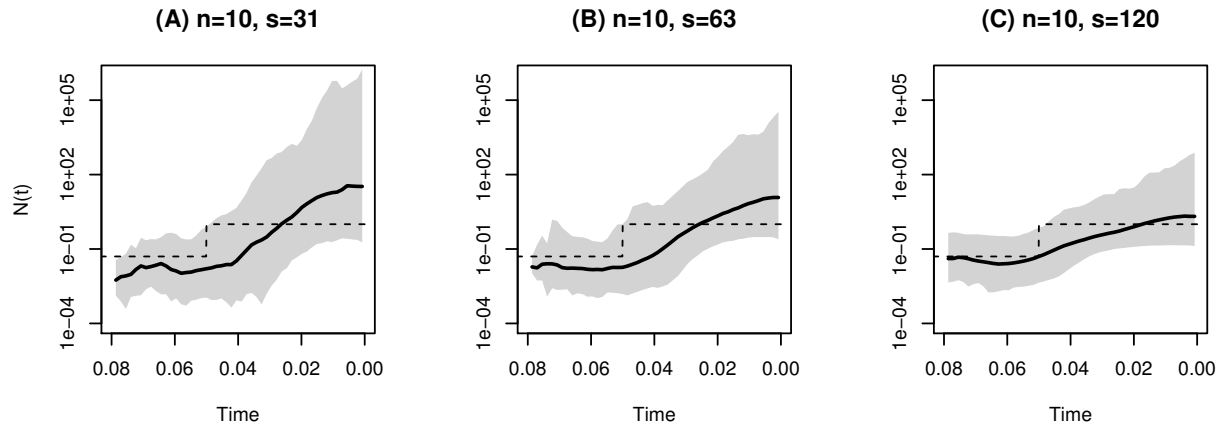


Figure 6: **Varying the number of segregating sites.** Posterior inference from simulated data of  $n = 10$  sequences under a population size trajectory with instantaneous growth (dashed lines).  $s$  is the number of segregating sites. Posterior medians are depicted as solid black lines and 95% Bayesian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

377 expect our method to better recover the truth. Figure 6 shows our results on simulated data under  
 378 a population size trajectory with instantaneous growth (Equation 13) of  $n = 10$  individuals with  
 379 31, 63 and 120 segregating sites. As expected, our method recovers the truth with higher precision  
 380 (MRW) and accuracy (SRE) when we increase the number of segregating sites. Increasing the  
 381 number of segregating sites may result in more constraints in the gene tree. For  $n = 10$ , there are  
 382 7936 possible ranked tree shapes, however for the datasets simulated with 31, 63 and 102 segregating  
 383 sites, there are only  $2582 \pm 32$ ,  $2670 \pm 34$  and  $556 \pm 7$  ranked tree shapes compatible with their  
 384 corresponding gene trees. These numbers were estimated by importance sampling (Cappello and  
 385 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

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



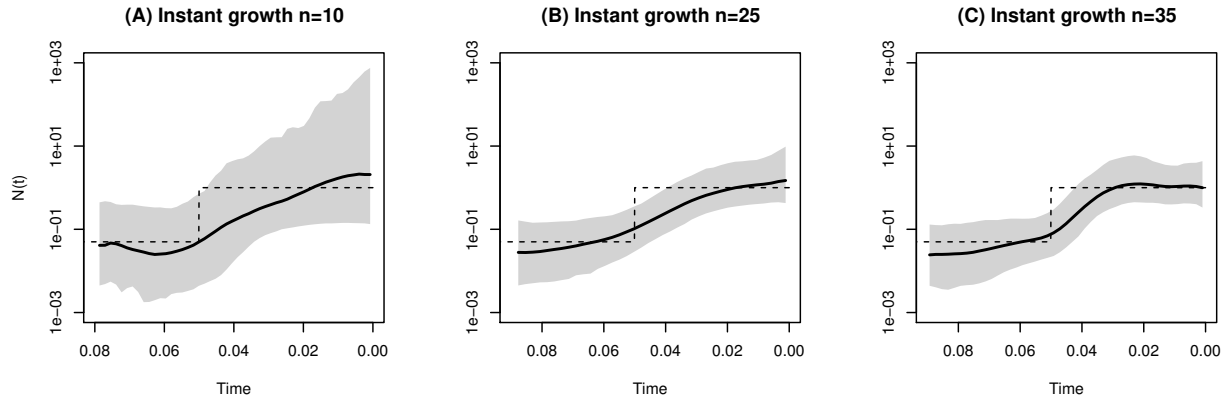


Figure 7: **Varying the number of samples under a population size trajectory with instantaneous growth.** Posterior inference from simulated data of  $n = 10, 25$  and  $35$  sequences under the population size trajectory with instantaneous growth. Shaded areas correspond to 95% credible intervals, solid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

387 with instantaneous growth with varying number of samples. We simulated datasets with  $n = 10,$   
388 25 and 35 samples with 215 expected number of segregating sites. Our results depicted in Figure  
389 7 show that our method performs better in terms of SRE and MRE when the number of samples  
390 increases. Similarly, precision (MRW) and accuracy (SRE) increases when inference is done from  
391 a larger number of independent datasets. Finally, Figure 8 shows our results from 1, 5 and 10  
392 datasets simulated from 1, 5 and 10 independent genealogies of 10 individuals with a population  
393 size trajectory of growth followed by a constant period (Equation 12). As expected, our method's  
394 performance substantially increases by increasing the number of genealogies.

### 395 3.2 Comparison to other methods

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

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

409 We compare the results of BESTT depicted in Figure 5 to those of BEAST (Drummond et al.,  
410 2005, 2012) in Figure 9. We note that results from BEAST are generated from 10 million iterations  
411 and thinned every 1000 iterations, while results from BESTT are generated from 200 thousand  
412 iterations.

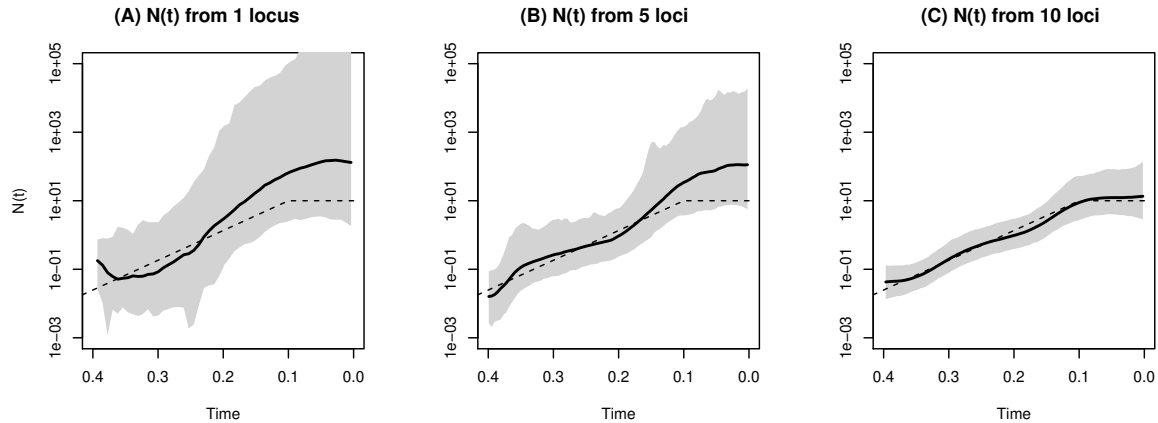


Figure 8: **Multiple independent datasets.** Posterior inference from simulated data of  $n = 10$  sequences under exponential followed by constant trajectory (eq. 12). **(A)** Inference from a single simulated dataset, **(B)** from 5 independently simulated datasets, and **(C)** from 10 independently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

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

Table 2: Performance comparison between BESTT and BEAST in simulations

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

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

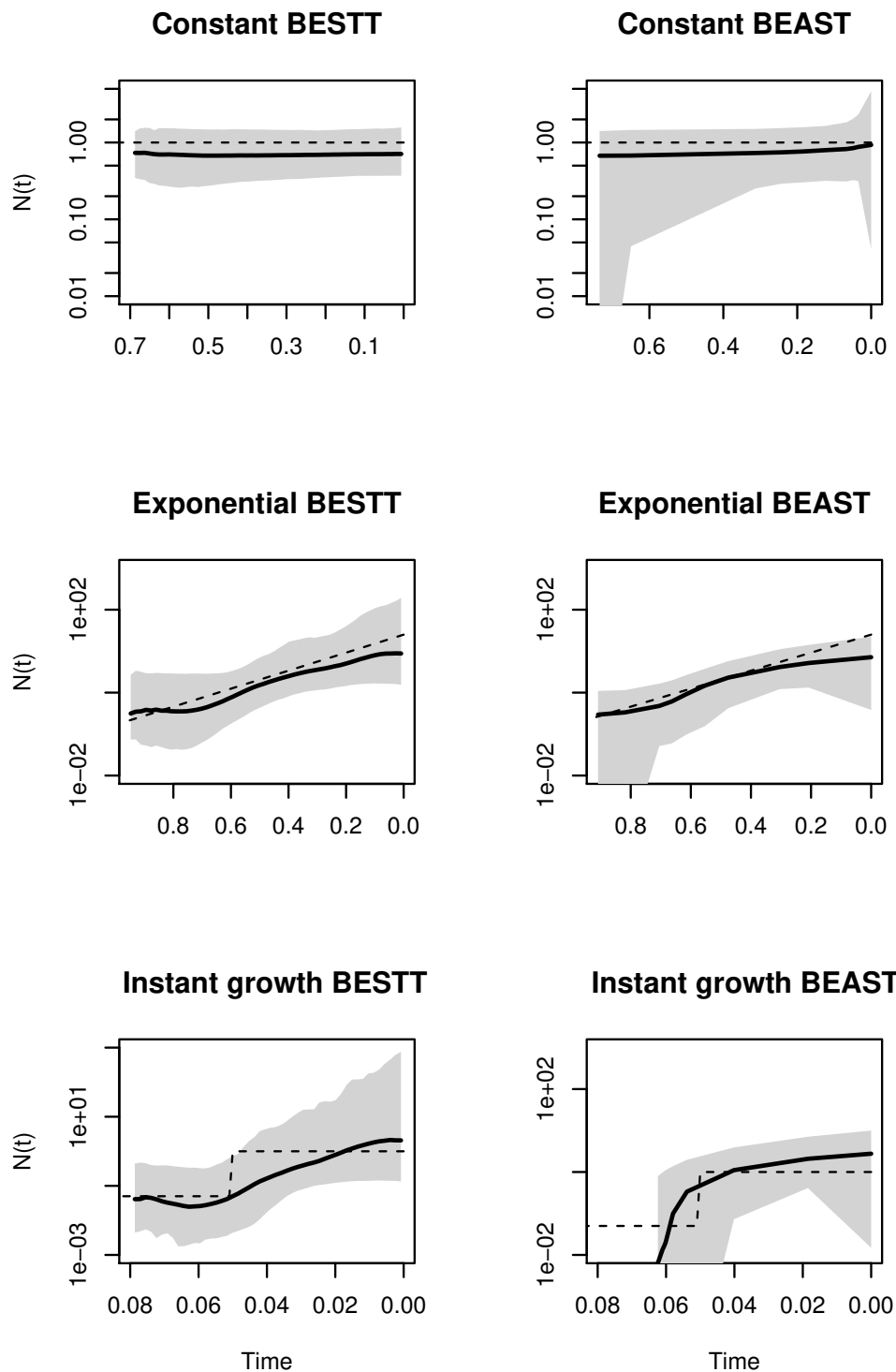


Figure 9: **BESTT and BEAST Comparison.** Posterior inference from simulated data of  $n = 10$  sequences under constant, exponential and instantaneous growth trajectories (rows) obtained from our method BESTT (first column) and BEAST (second column). Shaded areas correspond to 95% credible intervals,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

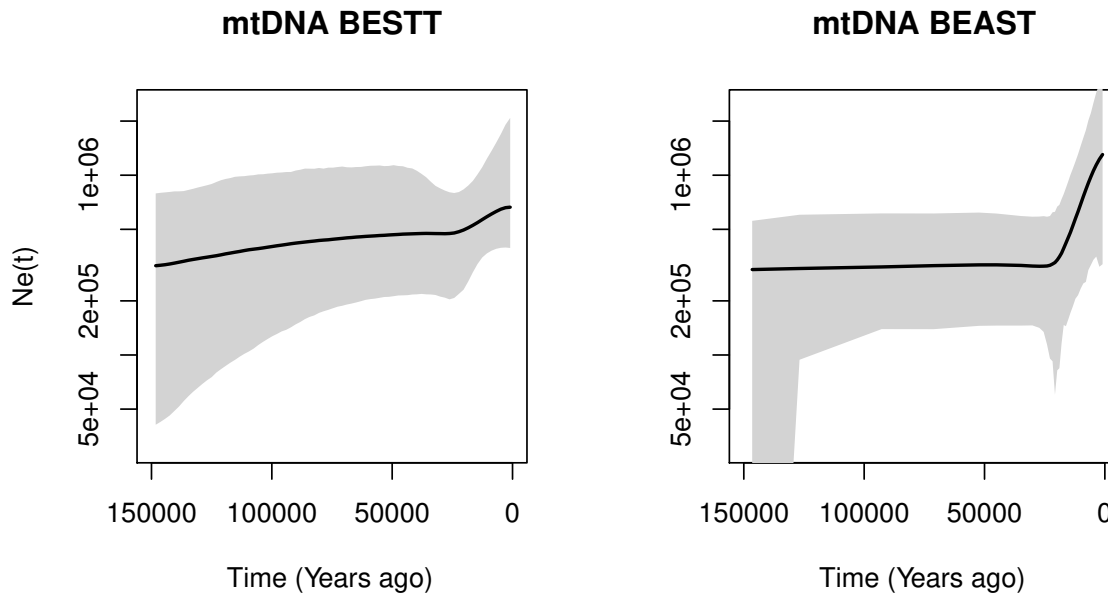


Figure 10: Posterior inference of female effective population size from 35 mtDNA samples from Yoruban individuals in the 1000 Genomes Project using our method BESTT (first plot) and the BEAST Extended Bayesian Skyline Plot (second plot). Posterior median curves are depicted as solid black lines and 95%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

#### 425 **4 Inferring human population demography from mtDNA**

426 We selected  $n = 35$  samples of mtDNA at random from 107 Yoruban individuals available from the  
427 1000 Genomes Project phase 3 (The 1000 Genomes Project Consortium, 2015). We retained the  
428 coding region: base pairs 576 – 16,024 according to the rCRS reference of Human Mitochondrial  
429 DNA (Anderson et al., 1981; Andrews et al., 1999) and removed 38 indels. Of the 260 polymor-  
430 phic sites, we retained 240 sites compatible with the infinite sites mutation model. The final file  
431 is available in <https://github.com/JuliaPalacios/phyloDYN>. To encode our data as 0s and 1s,  
432 we use the inferred root sequence **RSRS** of Behar et al. (2012) to define the ancestral type at  
433 each site. To rescale our results in units of years, we assumed a mutation rate per site per year of  
434  $1.3 \times 10^{-8}$  (Rebolledo-Jaramillo et al., 2014). We compare our results with the Extended Bayesian  
435 Skyline method (Drummond et al., 2012) implemented in BEAST in Figure 10. When applying  
436 BEAST, we assumed the Jukes-Cantor mutation model. Both methods detect an inflection point  
437 around 20kya followed by exponential growth. The mean time to the most recent ancestor (TM-  
438 RCA) inferred for these YRI mtDNA samples with BESTT is around 170kya with a 95% BCI of  
439 (142868, 207455), while the mean TMRCA inferred with BEAST is around 160kya with a 95% BCI  
440 of (133239, 196900). In Appendix B, we include two more comparisons of BESTT and BEAST.

## 441 5 Discussion

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

451 In this manuscript we show that a lower resolution coalescent model on genealogies, the “Tajima’s  
452 coalescent”, can be used as an alternative to the standard Kingman coalescent model. In particu-  
453 lar, we show that the Tajima coalescent model provides a feasible alternative that integrates over a  
454 smaller state space than the standard Kingman model. The main advantage in Tajima’s coalescent  
455 is to model the ranked tree topology as opposed to the fully labeled tree topology as in Kingman’s  
456 coalescent.

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

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

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

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

486 trajectories should improve mixing and convergence.

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

## 490 Acknowledgements

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

## 498 References

- 499 Aberer, A. J., Stamatakis, A., and Ronquist, F. (2016). An efficient independence sampler for up-  
500 dating branches in bayesian markov chain monte carlo sampling of phylogenetic trees. Systematic  
501 Biology, 65(1):161.
- 502 Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon,  
503 I. C., Nierlich, D. P., Roe, B. A., Sanger, F., et al. (1981). Sequence and organization of the  
504 human mitochondrial genome. Nature, 290(5806):457.
- 505 Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N.  
506 (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial  
507 dna. Nature genetics, 23(2):147.
- 508 Behar, D. M., Van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N. M., Kivisild, T.,  
509 Torroni, A., and Villems, R. (2012). A “copernican” reassessment of the human mitochondrial  
510 dna tree from its root. The American Journal of Human Genetics, 90(4):675–684.
- 511 Cappello, L. and Palacios, J. A. (2019). Sequential importance sampling for multi-resolution  
512 kingman-tajima coalescent counting. arXiv preprint arXiv:1902.05527.
- 513 Disanto, F. and Wiehe, T. (2013). Exact enumeration of cherries and pitchforks in ranked trees  
514 under the coalescent model. Mathematical Biosciences, 242(2):195 – 200.
- 515 Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. Annual  
516 Review of Genetics, 29(1):401–421.
- 517 Drummond, A. and Rodrigo, A. (2000). Reconstructing genealogies of serial samples under the  
518 assumption of a molecular clock using serial-sample UPGMA. Molecular Biology and Evolution,  
519 17(12):1807–1815.

- 520 Drummond, A., Suchard, M., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with  
521 BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29:1969–1973.
- 522 Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent infer-  
523 ence of past population dynamics from molecular sequences. Molecular Biology and Evolution,  
524 22(5):1185–1192.
- 525 Gill, M., Lemey, P., Faria, N., Rambaut, A., Shapiro, B., and Suchard, M. (2013). Improving  
526 bayesian population dynamics inference: A coalescent-based model for multiple loci. Molecular  
527 Biology and Evolution, 30:713–724.
- 528 Griffiths, R. and Tavaré, S. (1994a). Simulating probability distributions in the coalescent.  
529 Theoretical Population Biology, 46(2):131 – 159.
- 530 Griffiths, R. and Tavaré, S. (1995). Unrooted genealogical tree probabilities in the infinitely-many-  
531 sites model. Mathematical Biosciences, 127(1):77 – 98.
- 532 Griffiths, R. C. and Tavaré, S. (1994b). Sampling theory for neutral alleles in a varying environ-  
533 nment. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences,  
534 344(1310):403–410.
- 535 Griffiths, R. C. and Tavaré, S. (1996). Monte Carlo inference methods in population genetics.  
536 Mathematical and Computer Modelling, 23(8-9):141–158.
- 537 Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees. Networks, 21(1):19–28.
- 538 Heled, J. and Drummond, A. (2008). Bayesian inference of population size history from multiple  
539 loci. BMC Evolutionary Biology, 8(1):1–289.
- 540 Hobolth, A., Uyenoyama, M. K., and Wiuf, C. (2008). Importance sampling for the infinite sites  
541 model. Statistical applications in genetics and molecular biology, 7.
- 542 Jukes, T. H. and Cantor, R. C. (1969). Evolution of protein molecules. In Mammalian Protein  
543 Metabolism, pages 21–132. Academic, New York.
- 544 Karcher, M. D., Palacios, J. A., Lan, S., and Minin, V. N. (2017). phylodyn: an r package for  
545 phylodynamic simulation and inference. Molecular Ecology Resources, 17(1):96–100.
- 546 Kingman, J. (1982a). The coalescent. Stochastic Processes and their Applications, 13(3):235–248.
- 547 Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations. In Koch,  
548 G. and Spizzichino, F., editors, Exchangeability in Probability and Statistics, pages 97–112.  
549 North-Holland, Amsterdam.
- 550 Kuhner, M. and Smith, L. (2007). Comparing likelihood and Bayesian coalescent estimation of  
551 population parameters. Genetics, 175(1):155–165.
- 552 Kuhner, M. K. (2006). LAMARC 2.0: maximum likelihood and bayesian estimation of population  
553 parameters. Bioinformatics, 22(6):768.
- 554 Kuhner, M. K., Yamato, J., and Felsenstein, J. (1998). Maximum likelihood estimation of popula-  
555 tion growth rates based on the coalescent. Genetics, 149(1):429–434.

- 556 Lan, S., Palacios, J. A., Karcher, M., Minin, V., and Shahbaba, B. (2014). An efficient Bayesian  
557 inference framework for coalescent-based nonparametric phylodynamics. CoRR, abs/1126456.
- 558 Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome  
559 sequences. Nature, 475(7357):493–496.
- 560 Markovtsova, L., Marjoram, P., and Tavaré, S. (2000). The age of a unique event polymorphism.  
561 Genetics, 156(1):401–409.
- 562 Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough  
563 skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and  
564 Evolution, 25(7):1459–1471.
- 565 Palacios, J. A. and Minin, V. N. (2013). Gaussian process-based Bayesian nonparametric inference  
566 of population trajectories from gene genealogies. Biometrics, 63:8–18.
- 567 Palacios, J. A., Wakeley, J., and Ramachandran, S. (2015). Bayesian nonparametric inference of  
568 population size changes from sequential genealogies. Genetics.
- 569 Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral  
570 population sizes using dna sequences from multiple loci. Genetics, 164(4):1645–1656.
- 571 Rebolledo-Jaramillo, B., Su, M. S.-W., Stoler, N., McElhoe, J. A., Dickins, B., Blankenberg, D.,  
572 Korneliussen, T. S., Chiaromonte, F., Nielsen, R., Holland, M. M., Paul, I. M., Nekrutenko, A.,  
573 and Makova, K. D. (2014). Maternal age effect and severe germ-line bottleneck in the inheritance  
574 of human mitochondrial dna. Proceedings of the National Academy of Sciences, 111(43):15474–  
575 15479.
- 576 Sainudiin, R., Stadler, T., and Véber, A. (2015). Finding the best resolution for the kingman-tajima  
577 coalescent: theory and applications. Journal of Mathematical Biology, pages 1–41.
- 578 Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from  
579 multiple genome sequences. Nature Genetics, 46(8):919–925.
- 580 Schliep, K. (2011). phangorn: phylogenetic analysis in R. Bioinformatics, 27(4):592–593.
- 581 Sheehan, S., Harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from  
582 multiple genomes: A sequentially Markov conditional sampling distribution approach. Genetics,  
583 194(3):647–662.
- 584 Slatkin, M. and Hudson, R. (1991). Pairwise comparisons of mitochondrial DNA sequences in  
585 stable and exponentially growing populations. Genetics, 129(2):555–562.
- 586 Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. Journal of the  
587 Royal Statistical Society: Series B (Statistical Methodology), 62(4):605–635.
- 588 The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation.  
589 Nature, 526:68–74.
- 590 Watterson, G. (1975). On the number of segregating sites in genetical models without recombina-  
591 tion. Theoretical Population Biology, 7(2):256–276.



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

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

## 597 6 Appendix A

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

---

**Algorithm 1** Transition proposals for ranked tree shapes

---

**Input:**  $\mathbf{F}_n$

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

1. Set  $\mathbf{F}_n = \mathbf{F}_n^*$ .
2. Sample with uniform discrete probability a coalescent event  $k$  from the set  $\{1, \dots, n-2\}$ . 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.  $q(\mathbf{F}_n^* | \mathbf{F}_n) = q_1 q_2$ ,  $q(\mathbf{F}_n | \mathbf{F}_n^*) = q_1 q'_2$ .
-

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

---

**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**
- 

## 606 7 Appendix B

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

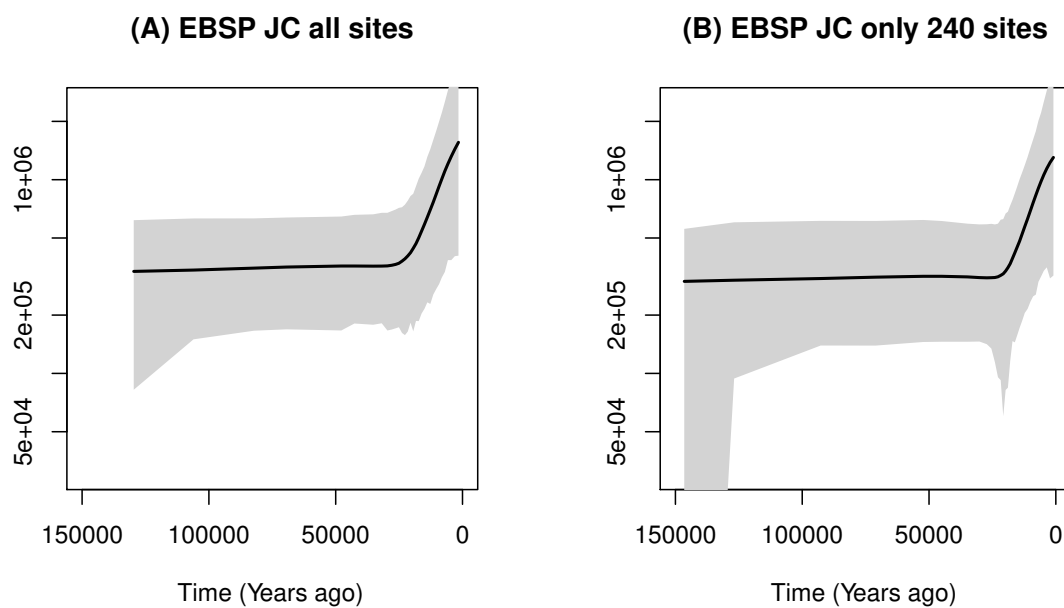


Figure 11: Posterior inference of female effective population size from 35 mtDNA samples from Yoruban individuals in the 1000 Genomes Project using BEAST EBSP (first plot) from all 15409 sites and the BEAST EBSP (second plot) from the 240 segregating sites retained. In both cases, the mutation model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

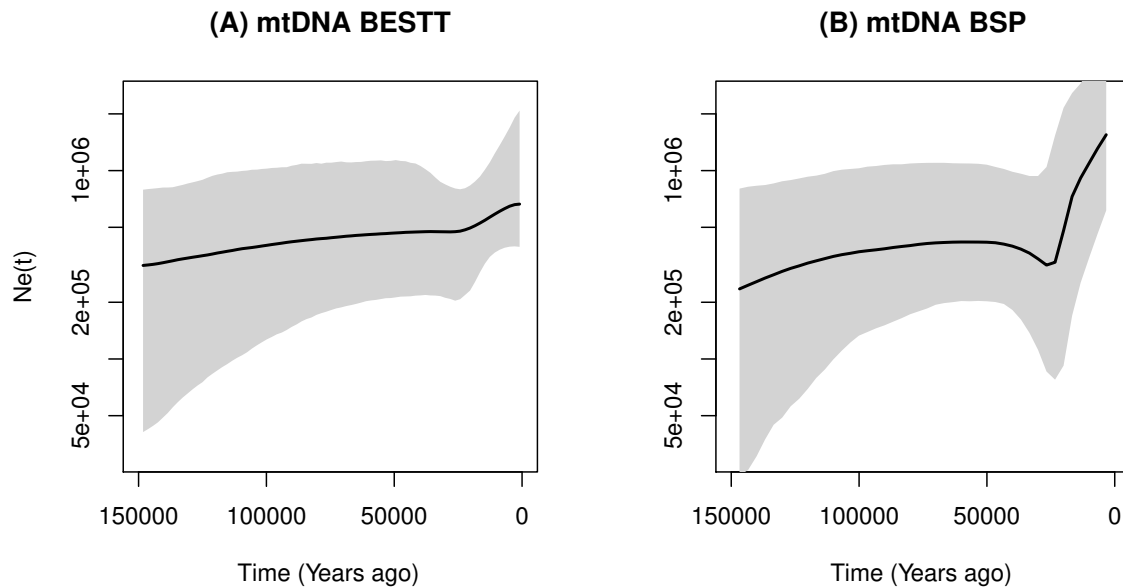


Figure 12: Posterior inference of female effective population size from 35 mtDNA samples from Yoruban individuals in the 1000 Genomes Project using our BESTT (first plot) from only 240 segregating sites and the BEAST BSP (second plot) from all the 15409 sites. Posterior median curves are depicted as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

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