Full Likelihood Inference from the Site Frequency Spectrum based on the Optimal Tree Resolution

Raazesh Sainudiin*, Amandine Véber †

August 26, 2017

*Department of Mathematics, Uppsala Universitet, Uppsala, Sweden

†CMAP, CNRS, École Polytechnique, Palaiseau, France

Running Head: Likelihood of Non-recombining Site Frequency Spectrum

Key Words: Importance sampler, semi-parametric estimation, optimal tree resolution, controlled Markov process on hidden genealogical trees

Corresponding Author:

Amandine Véber

CMAP - Ecole Polytechnique

route de Saclay

91128 Palaiseau Cedex

FRANCE

amandine.veber@cmap.polytechnique.fr

Abstract

We develop a novel importance sampler to compute the full likelihood function of a demographic or structural scenario given the site frequency spectrum (SFS) at a locus free of intra-locus recombination. This sampler, instead of representing the hidden genealogy of a sample of individuals by a labelled binary tree, uses the minimal level of information about such a tree that is needed for the likelihood of the SFS and thus takes advantage of the huge reduction in the size of the state space that needs to be integrated. We assume that the population may have demographically changed and may be non-panmictically structured or under the influence of some selection pressure, as reflected by the branch lengths and the topology of the genealogical tree of the sample, respectively. We also assume that mutations conform to the infinitely-many-sites model. We achieve this by a controlled Markov process that generates 'particles' in the hidden space of SFS histories which are compatible with the observed SFS. We use Aldous' Beta-splitting model for a one parameter family of prior distributions over genealogical topologies or shapes (including that of the Kingman coalescent) and allow the branch lengths or epoch times to have a parametric family of priors specified by a model of demography (including exponential growth and bottleneck models). Assuming independence across unlinked loci, we can obtain likelihood estimates of both the tree shape parameter β and the parameters dictating the distribution of epoch times according to a demographic model. Simulations studies are conducted to demonstrate the capabilities of the approach with publicly available code.

INTRODUCTION

Demographic inference based on genetic data has been a major challenge in the last two decades. Many methods and algorithms have been developed to turn the genetic diversity observed in a sample of individuals into reliable estimates of population structure or demography. Most of them consist in decomposing the question into (i) computing a weight or likelihood function for the parameters of interest under a given genealogical tree of the sample, and (ii) aggregating these weights by averaging over the set of all possible genealogies to account for the fact that they are hidden in practice. However, even for moderately large sample sizes ($n \geq 5$ or 10), the size of the state space of the full genealogy is huge and such an averaging cannot be performed in an exact way. Instead, one resorts to exploring the set of labelled trees as exhaustively as possible, for e.g. via Monte Carlo methods, but the associated computational cost grows extremely quickly with the sample size. In this work, we exploit the fact that the distribution of the *Site Frequency Spectrum* (SFS) depends on a coarser description of the genealogical tree of the sample than the classical leaf-labelled binary tree representation, which significantly decreases the size of the space to explore when computing (approximate) likelihoods.

Let us consider a locus which is free of intra-locus recombination, at which mutation occurs at rate θ. We sample n individuals at present and look at how mutations are shared among them. Under the infinitely-many-sites model, these Poissonian mutations give rise to a site frequency spectrum $S = (S_1, \ldots, S_{n-1})$, which reports how many mutations are carried by each given number of individuals in the sample:

$$S_j = \text{ number of mutations carried by } j \text{ individuals.} \tag{1}$$

The site frequency spectrum is used routinely in population genetics inference. See Gattepaille *et al.* (2013) and references therein for a recent review.

SFS carries considerable information about the underlying hidden genealogical tree with mutations upon it. In this paper, we obtain the likelihood of a demographic model and a

parameter specifying a measure of tree balance from the observed SFS. To do so, we develop a sampler which essentially takes a particular SFS S and produces a *tree topology matrix* F, a *mutation pattern matrix* M on the tree, and a vector of *epoch times* $T = (T_2, \ldots, T_n)$, such that F, M and T are compatible with the given SFS S, i.e. $\mathbb{P}(S \mid F, M, T) > 0$. Here, the epoch time T_k is the amount of time during which the sample has exactly k ancestors.

The main strengths of our approach reside in the following points. First, we use the minimal tree resolution on which the law of the SFS depends, in the sense of Sainudiin *et al.* (2015). More precisely, instead of the set of all leaf-labelled binary tree topologies (on which the Kingman coalescent is defined), we work with the set of all binary tree shapes (also called the *unvintaged and sized* tree resolution) encoded by our tree topology matrix $(F_{k,j})_{k,j}$ which only tracks the number of edges in the k-th epoch of the tree that subtend j leaves through $F_{k,j}$. See below for a more precise definition. This drastically reduces the state space of our sampler. Indeed, by Proposition 12 in Sainudiin *et al.* (2015), to each binary tree shape correspond of the order of $n!/2^{n-1}$ (at the very least) leaf-labelled binary tree topologies; as an illustration, note that $n!/2^{n-1}$ is approximately 7100 for $n = 10$, and 4.6×10^{12} for $n = 20$. Furthermore, only a subset of these shapes is compatible with the SFS and our sampler is able to produce only compatible tree shapes, each such shape having a positive probability of being produced (that is, no compatible shape is potentially missed by the sampler). The significant reduction of the state space of tree topologies enables us to consider a number of independent non-recombining loci which can be as high as a few thousands to combine the information one can derive from each of their SFS.

Second, the sampler needs to be informed of population-level processes equally affecting the genealogies at all independent loci only through (i) their tree balance via the parameter $\beta \in (-2, \infty)$ in Aldous' Beta-splitting model (Aldous 2001), introduced below to account for the effects of population structure, and (ii) their branch lengths via a vector of (inverses of) *a priori* mean epoch times which depend on the demographic scenario experienced by the population (but not on the SFS itself). In fact, one major difference between the signature

left by population structure and that left by fluctuating population sizes lies in the balance of the genealogical tree topologies, which has been barely explored even in the classical models. This characteristic balance of the genealogies influences the observed mutation pattern and could thus be at the basis of a straight-forward test for population structure affecting all loci or nonneutral evolution affecting only loci under selection or linked to a recently selected site. The importance sampler and likelihood procedure developed in this paper allow not only to test for deviations from panmixia and neutrality (corresponding to $\beta = 0$ in the sequel), but also to infer the most likely balance parameter β corresponding to a given SFS, or set of SFSs from independent loci representing processes that affect the whole population.

As described above, the sampling of a 'particle' (F, M, T) uses the Beta-splitting model of tree balance in the topology matrix F, and only requires some *a priori* demographic laws for epoch times in T. The expected epoch times under these demographic laws can be obtained either analytically, or by an easy round of simulations from *any* standard demographic model, including parametric models, such as exponential growth or bottleneck, or semi-parametric models involving the class of piecewise constant or exponential functions, for example.

A rather large panel of methods already exist to infer demographic parameters from the observed SFS. The *Poisson Random Field* approach (Sawyer and Hartl 1992; Nielsen 2000; Gutenkunst *et al.* 2009) considers a series of independent SNPs in a sample of size n. Assuming the infinitely-many-sites mutation model with a very low mutation rate, the distribution of the number of mutations (or derived alleles) carried by $k \in \{1, \ldots, n-1\}$ individuals is either approximated by a Poisson distribution whose parameter is given by the ratio of the average length of all the edges in the genealogical tree that subtend k leaves to the average total length of the tree; or it is described as a Poisson random variable whose parameter is given by the probability that k out of n individuals are sampled within the current population carrying the derived allele. The average length of edges subtending a given number of sampled individuals is usually obtained by simulation as in Nielsen (2000), while the sampling probability is obtained by solving a Wright-Fisher diffusion with

selection as in Gutenkunst *et al.* (2009). Our procedure generalizes this approach in that its likelihood is not restricted to a Poisson decomposition of a single global genealogical expectation of the Kingman coalescent based models across all loci, and we can handle full SFS data by constructing locus-specific particles whose genealogical and mutational histories are compatible with the SFS at each one of several independent loci which are free of intra-locus recombination. In contrast, the Poisson random field approaches impose that each locus should be a single segregating site that is at an infinite recombinational distance from all other segregating sites, and thereby miss the shared genealogical signal at linked sites. Assuming that the mutation rate is very small so that we see at most one mutation per locus, the recombinational distance between adjacent loci is very large, and the tree balance parameter β is 0 in order to enforce the Kingman coalescent prior on tree topologies, our sampler is essentially equivalent to the Poisson Random Field approach of Nielsen (2000), with the notable difference that the probability of seeing a mutation carried by k individuals is now computed from the true probability of the placement of the mutation *conditionally on the tree topology*, for every particle generated by the sampler for each locus with possibly more than one segregating sites.

In Bunnefeld *et al.* (2015), a method based on the probability generating function of the branch lengths in the genealogies is developed to extract the signal in SFS from linked sites and applied to detect the occurrence of a bottleneck in the history of the population, relying again on the Kingman coalescent model. Despite the use of unlabelled tree shapes and other clever tricks to take advantage of the symmetries of the problem, deriving the probability of a given SFS requires the sum over a very large set of mutation placements on the tree which are compatible with the SFS, a generally unfeasible computation whenever $n > 5$.

Skyline plots form another family of inference methods for demographic history, as reviewed in Ho and Shapiro (2011). These nonparametric methods rest on the assumption that there is not much variability in the SFS-compatible reconstructed tree on which the

estimation of the local harmonic means of effective population size is based (c.f., Pybus *et al.* (2000)). However, this will typically not be the case when the individual mutation rate is low and the SFS contains only a few mutations. Heled and Drummond (2008) extend the method to several unlinked loci, enhancing the demographic signal captured by the reconstructed trees at the different loci.

All these approaches assume independent loci free of intra-locus recombination, which may be sensible if we consider short loci far enough from each other in the genome. Following the improvement of the accessibility of whole-genome data and of the mathematical modelling of linkage, different methods focusing on large stretches of recombining DNA have been set up and used to reconstruct population histories. For instance, Harris and Nielsen (2013) study the set of distances between neighbouring SNPs within a long sequence of genome from a sample of two individuals. They derive an approximate formula for the distribution of the typical length of a tract of identity by state (IBS) using the Sequentially Markov Coalescent (SMC) of McVean and Cardin (2005) and the related SMC′ model of Marjoram and Wall (2006). Assuming these pieces of IBS sequences are nearly independent, they use a composite likelihood approach to infer the parameters in a model incorporating population size changes, and divergence and admixture events between sub-populations. The same approach is used by Barton *et al.* (2013) to reconstruct the lineage diffusion coefficient and the neighbourhood size in a spatially structured neutral population. The SMC or SMC′ approximation is also a key ingredient in the conditional haplotype sampling distribution developed by Steinrücken *et al.* (2016) for population scenarios comprising discrete sub-populations related through migration and with potentially varying effective population sizes (described by a given class of functions, such as piecewise linear or piecewise exponential). Palacios *et al.* (2015) model the effective population size of a population as the exponential of a Gaussian process. Assuming that the local genealogies follow the SMC′ model, the pattern of diversity observed in the sequence data is used to reconstruct the fluctuations of the effective population size in a Bayesian nonparametric way. Note that the choice made there to

encode the tree topologies at the sufficient resolution of the *ranked tree shapes* (Tajima 1983; Sainudiin *et al.* 2015) considerably enhances the exploration of this component of the state space during the MCMC step, a point which is pushed further by Palacios *et al.* (2017). Such methods explicitly try to model recombination within a locus, involving a very large hidden space of compatible histories when compared to recombination-free histories. Our approach generalizes the Poisson random field methods and can be complementary to methods that model recombination explicitly when applied to recombinationally distant blocks of contiguous sites which are devoid of any signal of recombination within the block. This assumption, along with our parametric model for tree balance, allows us to extract information from SFS in a locus-specific manner across thousands of loci that are recombinationally independent for demographic inference and outlier detection.

To overcome the difficulty of computing analytical (or even approximate) likelihoods in potentially complex population models, a simple simulation-intensive approach known as Approximate Bayesian Computation or ABC (Beaumont *et al.* 2002) is now routinely used in a wealth of studies. Recently Boitard *et al.* (2016) demonstrate through simulations that considering the joint information provided by SFS and linkage disequilibrium improves the accuracy of parameter reconstruction when compared to a method based only on SFS or LD. An ABC methodology is used by Peter *et al.* (2010) to distinguish different demographic and structural scenarios using a variety of statistics of microsatellite data. A significant disadvantage of any ABC method is the lack of locus-specific likelihood for the SFS itself, the basic quantity that the method tries to circumnavigate from computing. Typically, the SFS data across multiple loci is reduced to the mean and variance of various *ad hoc* summary statistics of the SFS across all the loci during the simulation-intensive approximations underlying ABC and thus no efforts are made to integrate over hidden genealogical histories that are compatible with the SFS at each locus. In contrast, our approach develops a controlled Markov process to obtain the likelihood of each SFS directly.

The methods reviewed here are well-suited for inferring demographic parameters that

affect all loci, since they use the combined information coming from the per-locus site frequency spectra generated by their common history. In a few instances, they have also been used to detect outlier loci potentially subject to natural selection (Nielsen 2005; Roux *et al.* 2012). However, when one wants to infer the locus-specific history provided by the SFS, to our knowledge there are no known methods analogous to the importance samplers available for inference from the full *binary incidence matrix* or BIM data (Fearnhead and Donnelly 2001; De Iorio and Griffiths 2004; Hobolth *et al.* 2008; Koskela *et al.* 2015; Kamm *et al.* 2016), except the naive importance sampler via controlled Markov chains in Sainudiin *et al.* (2011). Here we propose an efficient importance sampler that uses very natural *a priori* information on the law of hidden genealogical histories to produce a triplet (F, M, T) at the minimally sufficient resolution of tree topology, mutation history and epoch times that are compatible with a given SFS. Since the hidden state space has been optimized, our method can cope with large numbers of samples and independent loci to obtain maximum likelihood estimates or MLEs of (i) demographic parameters under models such as exponential growth or bottleneck or other semi-parametric models and (ii) the Beta-splitting parameter for tree balance (including the Kingman coalescent) which affect the entire population. Furthermore, since the particle systems in the hidden space are constructed locus-specifically, the likelihood and MLEs at each locus can be used directly for outlier detection.

The proof-of-concept code for the sampler and the likelihood procedure is publicly available at Sainudiin and Véber (2017). The detailed pseudo-code can be found in Section 1 of the Supplementary Material.

THE SAMPLER

In all that follows, we assume that the genealogical tree relating a sample of n individuals is always binary. A coalescence event therefore corresponds to the number of ancestral lineages decreasing from some $k \in \{2, \ldots, n\}$ to $k-1$. We call epoch k the interval of time in the

past during which the sample has k distinct ancestors, and write T_k for the duration of this epoch. In other words,

$$T_k = \text{ amount of time during which the tree has } k \text{ edges.} \tag{2}$$

[Figure 1 about here.]

Mutations are assumed to occur at some per-lineage rate $\theta > 0$ along the branches of the tree, and to conform to the infinitely-many-sites model. No recombination happens within the locus considered. See Figure 1 for an example of mutations on a stretch of nine sites which are completely linked.

Instead of integrating over the full history of leaf-labelled coalescent trees with mutations (as shown on the left side of Figure 1), our main idea here is to work with a much smaller space of topology matrices encoding the number of edges in epoch k that subtend j leaves, and mutation matrices recording the number of mutations that fall on such edges (as shown on the right side of Figure 1 by F and M, respectively). Explicitly, for every $k \in \{2, \ldots, n\}$ and $j \in \{1, \ldots, n-1\}$,

$$F(k, j) = \text{number of edges in epoch } k \text{ subtending } j \text{ leaves,} \tag{3}$$

while

$$\begin{aligned} F(k, 0) &= \text{size of the largest edge created by the split between epochs } k-1 \text{ and } k, \\ F(k, n) &= \text{size of the edge split between epochs } k-1 \text{ and } k. \end{aligned} \tag{4}$$

where the size of an edge corresponds to the number of leaves (or sampled individuals) that it subtends. The rows and columns of F thus range in $2, 3, \ldots, n$ and $0, 1, \ldots, n$, respectively, and necessarily $F(k, j) = 0$ if $j \in \{n-k+2, \ldots, n-1\}$. Observe that the topology matrix F encodes more information than the f-sequence of (Sainudiin *et al.* 2011), the minimal sufficient topological statistic of the labelled genealogical tree of the sample (in addition to epoch times), that is necessary for the prescription of multinomial-Poisson probabilities for

SFS (see Eqs. (6) and (7)). The extra information in F given by $F(k,0)$ and $F(k,n)$ can in fact be derived from the other matrix entries, but are recorded in the two extra columns to ease the computations based on the F-matrix. Indeed, these coefficients are exactly what is needed to prescribe the topology probabilities dictating the tree shape under Aldous' Beta-splitting model (Aldous 2001), our simple parametric model of various phenomena affecting tree shape.

Similarly,

$$M(k,j) = \text{number of mutations carried by one of the } F(k,j) \text{ edges in epoch } k \text{ which subtend } j \text{ leaves.} \tag{5}$$

Therefore, the rows and columns of M range in $2,3,\ldots,n$ and $1,\ldots,n-1$, respectively. We use standard notation for the sub-matrix of a matrix: $F(a:b,c:d)$ is the sub-matrix of F made of rows $a,a+1,\ldots,b$ and columns $c,c+1,\ldots,d$.

The product of the epoch time row vector T and the matrix $F(2:n,1:n-1)$ gives the total length of the edges in the tree that subtend $1,2,\ldots,n-1$ leaves as follows:

$$\begin{aligned} L &= (L_1,L_2,\ldots,L_{n-1}) = T\times F(2:n,1:n-1) \\ &= \left(\sum_{k=2}^{n} T_k F(k,1), \sum_{k=2}^{n-1} T_k F(k,2), \ldots, \sum_{k=2}^{2} T_k F(k,n-1)\right) . \end{aligned} \tag{6}$$

As shown in Proposition 5 of Sainudiin *et al.* (2011), the probability of the SFS only depends on the coalescent tree through L: Conditionally on L,

$$\mathbb{P}\big[S=(s_1,\ldots,s_{n-1}) \,\big|\, L\big] = e^{-\theta(L_1+\cdots+L_{n-1})} \prod_{j=1}^{n-1} \frac{(\theta L_j)^{s_j}}{s_j!}. \tag{7}$$

Thus, we only need to consider topology matrices and epoch time vectors to compute the likelihood of the observed SFS. However, in the incremental construction of F and T by our sampler we use M to take the partially constructed SFS history into account while ensuring that the fully reconstructed SFS history remains consistent with the observed SFS.

Next, we describe the sampler's *a priori* laws for the tree topology and the epoch times.

***A Priori* Laws for the Topology.** Our sampler uses a modification of the Beta-splitting model of Aldous (2001) which specifies the order of the splits. Aldous' Beta-splitting model is a one-parameter family of random cladograms which has the advantage of containing several of the most classical null models for tree shapes used in phylogeny reconstruction (Mooers and Heard (1997)), such as the *equal-rates-Markov* model (i.e., the random topology of the Kingman coalescent), the *proportional-to-distinguishable-arrangements* model or the *equiprobable-types* model. More precisely, for a given choice of $\beta \in (-2, \infty)$, if an edge subtending b leaves (i.e., ancestral to b individuals in the sample) is split into two edges, then the probability that the two daughter edges subtend x and $b-x$ leaves is given by

$$\lambda_{b,x} = \begin{cases} 2a_b^{-1}\binom{b}{x}\int_0^1 u^{x+\beta}(1-u)^{b-x+\beta}du & \text{if } b/2 < x \leq b-1, \\ a_b^{-1}\binom{b}{x}\int_0^1 u^{x+\beta}(1-u)^{b-x+\beta}du & \text{if } x = b/2 \quad (\text{when } b \text{ is even}), \end{cases} \tag{8}$$

where a_b is the normalizing factor

$$a_b = \int_0^1 (1-u^b-(1-u)^b)u^\beta(1-u)^\beta du.$$

The particular case $\beta = 0$ corresponds to the topology of the Kingman coalescent. Choosing β close to -2 gives rise to comb-like trees, while $\beta \gg 1$ produces highly balanced tree topologies (Aldous 2001). Thus the Beta-splitting model gives us a one-parameter family spanning the whole range of possible tree balances.

Using (8), we can define the probability of producing a given topology F under our incremental Beta-splitting model. To simplify the notation, for every epoch $k \in \{2, \ldots, n\}$, we write m_k for the size (or number of leaves it subtends) of the edge split between epochs $k-1$ and k, and ℓ_k for the size of the largest edge created during this split. Recalling (4), we have $m_k = F(k, n)$ and $\ell_k = F(k, 0)$.

We proceed by going from the root towards the leaves of the tree. First, the edge chosen to break at the beginning of epoch $k \geq 2$ has size m with probability $F(k-1, m) * (m-1)/(n-k+1)$ (observe that $F(1, \cdot) = (0, \ldots, 0, 1)$, since epoch 1 formally corresponds to

the period during which there is only one ancestor to the whole n-sample). The law of this random choice is arbitrary since it is not specified in Aldous' *stickbreaking* construction (Aldous (2001)); it corresponds to the law of the choice of the block to be split at the beginning of the k-th epoch in the *forwards-in-time* unlabelled Kingman coalescent. Second, the split within the chosen block is given by the probability $\lambda_{m,\cdot}$ described in (8). Thus, we obtain that under the law $\mathbb{P}_\beta$ of the incremental Beta-splitting model with parameter $\beta \in (-2, \infty)$, the probability of a given topology F is given by

$$\mathbb{P}_\beta(F) = \prod_{k=2}^{n} \left(\frac{F(k-1, m_k)(m_k - 1)}{n - k + 1} \lambda_{m_k, \ell_k} \right). \tag{9}$$

In particular, if $\beta = 0$, then

$$\lambda_{m,\ell} = \frac{2 - \mathbf{1}_{\{\ell = m/2\}}}{m - 1},$$

and if $\neg(F)$ denotes the number of splits such that $\ell_k \neq m_k/2$, we indeed recover the probability

$$\mathbb{P}_0(F) = \frac{2^{\neg(F)}}{(n-1)!} \prod_{k=2}^{n} F(k-1, m_k) \tag{10}$$

of the unvintaged and sized Kingman coalescent (see Proposition 11 in Sainudiin *et al.* (2015)).

***A Priori* Laws for the Epoch Times.** The epoch time component of our sampler is initialized with a sample from a vector $T^0 = (T_2^0, \ldots, T_n^0)$ of $n-1$ independent exponential random variables with respective parameters A_k. The mean A_k^{-1} of T_k^0 is taken to be the average length of the k-th epoch in the scenario whose likelihood we want to compute, independently of the observed SFS. For example, if we assume that the genealogy underlying the observed SFS conforms to the Kingman coalescent with effective population size N_0, then

$$\mathbb{E}\left[T_k^0\right] = \frac{2N_0}{k(k-1)}, \qquad 2 \leq k \leq n.$$

Other examples of parametric models are given in Section 3 of the Supplementary Material. The *a priori* rate vector $(A_k)_{2 \leq k \leq n}$ is thus another input of the sampler. As mentioned in

the introduction, its components can either be computed analytically in simple scenarios, or can be estimated by a round of simulations (without conditioning on the observed SFS).

The choice of exponential *a priori* laws for the epoch times is motivated by the following property of conditioned Gamma random variables: If T follows a Gamma distribution with parameters k, λ, denoted by $\mathcal{G}(k, \lambda)$ and with density

$$\frac{\lambda^k}{\Gamma(k)} t^{k-1} e^{-\lambda t} \mathbf{1}_{\{t>0\}} \tag{11}$$

(where Γ is the Gamma function), then the law of T conditional on $\text{Poisson}(\theta T) = m$ is again a Gamma distribution with parameters $k + m, \lambda + \theta$. Indeed, we have

$$\begin{aligned}
\mathbb{P}\big[\text{Poisson}(\theta T) = m\big] &= \frac{\lambda^k}{\Gamma(k)} \int_0^\infty t^{k-1} e^{-\lambda t} \mathbb{P}\big[\text{Poisson}(\theta t) = m\big] \mathrm{d}t \\
&= \frac{\lambda^k}{\Gamma(k)} \int_0^\infty t^{k-1} e^{-\lambda t} \frac{e^{-\theta t}(\theta t)^m}{m!} \, \mathrm{d}t \\
&= \frac{\lambda^k \theta^m}{\Gamma(k) m!} \int_0^\infty t^{k+m-1} e^{-(\lambda+\theta)t} \mathrm{d}t \\
&= \frac{\lambda^k \theta^m \Gamma(k+m)}{\Gamma(k)(m!)(\lambda+\theta)^{k+m}},
\end{aligned}$$

and so the density of T conditional on $\text{Poisson}(\theta T) = m$ is equal to

$$\frac{1}{\mathbb{P}\big[\text{Poisson}(\theta T) = m\big]} \frac{\lambda^k}{\Gamma(k)} t^{k-1} e^{-\lambda t} \frac{e^{-\theta t}(\theta t)^m}{m!} = \frac{(\lambda+\theta)^{k+m}}{\Gamma(k+m)} t^{k+m-1} e^{-(\lambda+\theta)t}$$

on $\mathbb{R}_+^*$. This is the density of the $\mathcal{G}(k + m, \lambda + \theta)$ distribution.

This property will be extensively used in the updating of the epoch times occurring during each step of the construction of a particle by the sampler.

Definition of a Particle. Write S for the observed SFS. For any fixed $A = (A_k)_{2 \le k \le n}$, $\beta \in (-2, \infty)$ and $\theta > 0$, our sampler produces samples from the support of

$$\mathbb{P}_{A,\beta,\theta}[(F, M, T) \mid S],$$

where $\mathbb{P}_{A,\beta,\theta}$ is the probability measure under which the tree topology follows the incremental Beta-splitting model with parameter β, the epoch times are exponentially distributed with

parameters A_k and mutations fall on the tree at rate θ. In particular, our sampler directly produces trees with mutations which are compatible with the observed SFS S, in the sense that no sampled SFS-history (F, M, T) satisfies

$$\mathbb{P}_{A,\beta,\theta}[(F, M, T) \mid S] = 0.$$

It is devised in a way which maximizes the exploration of the state space of trees with mutations that are compatible with the observed SFS S, while biasing the sampling according to the desired balance β and epoch times dictated by A.

The probability that the sampler produces an SFS history (F, M, T) is

$$q_{A,\beta,\theta}[(F, M, T) \mid S] \propto w w_1 w_2, \tag{12}$$

where w, w_1 and w_2 are the proposal weights corresponding to F, M and T, respectively. A *particle* refers to such an S-compatible SFS history and its proposal weights, i.e., $[(F, M, T), (w, w_1, w_2)]$. Importance sampling estimators are based on a collection of such particles and thus the most basic task of the sampler is the probabilistic construction of a particle. Below we provide an overview of how a particle is constructed incrementally in three stages. The detailed pseudo-code can be found in Section 1 of the Supplementary Material.

The Sampler. The construction of a particle (F, M, T) is an incremental process. It starts from a topology matrix F and a mutation matrix M whose entries are all zero, a vector of epoch times drawn from the independent exponential *a priori* distributions with rates given by the vector A, and the proposal weights w, w_1, w_2 are all set to 1. First, we use S_{n-1}, the number of mutations carried by exactly $n-1$ individuals in the sample, to force the presence of an edge subtending $n-1$ leaves if $S_{n-1} > 0$. We distribute the S_{n-1} mutations on the appropriate edges and update the corresponding epoch times. Then, we update the result of this partial construction by distributing the S_{n-2} mutations carried by $n-2$ individuals on the edges subtending $n-2$ leaves (which are forced to exist if $S_{n-2} > 0$) and by resampling the corresponding epoch times conditionally on the partial mutation pattern. We then use

the mutations carried by $n-3$ individuals to update the new value of (F, M, T) by possibly inserting some edges subtending $n-3$ leaves, and so on. Hence, the j-th step of this algorithm considers mutations carried by $n-j$ individuals and leads to the potential insertion of some $(n-j)$-edges (and only such edges) in the partial topology.

More precisely, each of these steps (say, the j-th one) goes through three stages:

(i) **Updating the topology:** We update the topology matrix F and the importance weight w obtained at the end of step $j-1$ based on whether $S_{n-j} > 0$ (and we should see at least one edge subtending $n-j$ leaves) or not. We start in epoch 2 and go down the partially constructed topology until there is a possibility for the creation of an $(n-j)$-edge. Suppose first that $n-j > n/2$, so that there may be at most one such edge in the tree, created by the splitting of a bigger edge. Because this 'parent' edge subtends $m > n-j$ leaves, the epoch at which it is split is already encoded in the partially constructed F matrix. At the moment when it splits, we also know that $n-j$ is the biggest size possible for the largest 'daughter' edge (otherwise the creation of a bigger edge would already be recorded). Thus, if $S_{n-j} > 0$ we force the creation of an $(n-j)$-edge at the epoch starting at the split. If $S_{n-j} = 0$ we have no constraints, and so we decide whether the split gives rise to an $(n-j)$-edge or not at random, using the Beta-splitting distribution $\lambda_{m,\cdot}$ with parameter β, *conditional* on the size of the largest daughter edge being at most $n-j$. We also multiply the importance weight w by the corresponding conditional probability. If the $(n-j)$-edge is indeed created, all the other edges present at the same time in the tree are necessarily smaller (recall that $n-j > n/2$). Thus, no split has yet been fixed in the remaining epochs until epoch n. As in the description of the incremental Beta-splitting model, we carry on going down the tree and decide at the beginning of each epoch k to split the $(n-j)$-edge with probability $(n-j-1)/(n-k+1)$, or to keep it with probability $(j+2-k)/(n-k+1)$, until the split occurs and the edge disappears from the later epochs (i.e., $F(k, n-j) = 0$ again). Note that the split probability is 1 when $k = j+2$, corresponding to the fact

that there can be no $(n-j)$-edges in epoch $k > j+1$. The weight w is updated accordingly.

When $n - j \leq n/2$, the procedure is similar but we have to take into account the information present in the partial topology resulting from the previous steps, which could force the creation of an $(n-j)$-edge independently of the presence of mutations carried by $n-j$ individuals in the sample (if the split of an m-edge and the subsequent creation of an $(m-(n-j))$-edge are already encoded in the partial topology, or if n is even and a split $(n/2, n/2)$ is the only remaining option for the first split). The different cases $1 \leq j < n/2$, $j = n/2$ and $n/2 < j \leq n-3$ are handled respectively by the procedures Sstep, Hstep and Lstep. Eventually, there is no randomness in the insertion of the edges subtending 2 or 1 leaves, which are carried out by the procedures Twostep and Onestep. All these procedures are listed in Section 1 of the Supplementary Material.

(*ii*) **Distributing the mutations:** The mutations carried by $n-j$ individuals are placed on the tree. We exploit the information given by the vector T produced by the previous steps (i.e., the information obtained from mutations carried by more than $n-j$ individuals) to give a weight to each epoch containing at least one edge subtending $n-j$ leaves. Then we distribute the S_{n-j} mutations in a multinomial way, using these weights. The simple idea behind this multinomial scheme is that if mutations fall on the tree like a Poisson point process with fixed rate, then conditionally on there being S_{n-j} mutations carried by $n-j$ individuals, they are independently and uniformly distributed over the total length of $(n-j)$-edges in the tree. We also update the current value of the importance weight w_1 by multiplying it by the probability of the mutation placement obtained. Of course, if $S_{n-j} = 0$ there is nothing to do, even if an $(n-j)$-edge exists.

(*iii*) **Updating the epoch times:** We only update the lengths of the epochs containing at

least one $(n-j)$-edge (since the distribution of mutations gives no new information on the epochs where there are no such edges). To this end, we use the stability property of the Gamma distributions expounded in the previous section: If $T \sim \mathcal{G}(m, \lambda)$, then T conditioned on the event $\{\text{Poisson}(\theta T) = s\}$ is a $\text{Gamma}(m+s, \lambda+\theta)$ random variable. Using the previous steps and the fact that an exponential distribution with rate A_k is also a $\mathcal{G}(1, A_k)$ distribution, for every epoch k we know that the current value of T_k is a random draw from a $\mathcal{G}(m_-, \lambda_-)$ distribution with

$$m_- = 1 + \sum_{l=n-j+1}^{n-1} M(k,l) \qquad \text{and} \qquad \lambda_- = A_k + \theta \sum_{l=n-j+1}^{n-1} F(k,l).$$

Thus, for every epoch k in which an $(n-j)$-edge was placed during the first stage, we draw a new value of T_k from a $\mathcal{G}(m_+, \lambda_+)$ distribution with

$$m_+ = 1 + \sum_{l=n-j}^{n-1} M(k,l) \qquad \text{and} \qquad \lambda_+ = A_k + \theta \sum_{l=n-j}^{n-1} F(k,l),$$

and multiply the importance weight w_2 corresponding to this component by the density of the Gamma variable at the value drawn.

MakeHistory listed as Function 1 in Section 1 of the Supplementary Material outputs the desired particle as a list $[F, M, T, w, w_1, w_2]$ when called with the following input arguments: the sample size n, the SFS S, the scaled mutation rate θ, the tree shape parameter β and the vector A of *a priori* rates for the epoch times. The presence or absence of mutations carried by each number of individuals is encoded in a control sequence $C = (C_1, \ldots, C_{n-1})$, constructed at the beginning of the procedure by setting $C_j = 1$ if $S_j > 0$, and $C_j = 0$ otherwise. The control value $C_j = 1$ forces the insertion of a first j-edge in the tree, after which C_j is set to 0 and the insertion of other j-edges remains possible but is not compulsory.

SIMULATIONS

We explored the properties of our sampler through simulations of two standard models of demography, the exponential growth model and the bottleneck model. Both are recalled and

their likelihood functions are given in Section 3 of the Supplementary Material. We also produced data with different β parameters in order to test whether our likelihood procedure was able to detect deviations from the Kingman (or equal rates) topology corresponding to $\beta = 0$. Recall that $\beta \to -2$ yields more and more imbalanced trees, while $\beta \to \infty$ gives rise to more and more balanced trees. As a sanity check, we also produced simulations with very large mutations rates (yielding of the order of 300 or more SNPs per locus for a sample size of $n = 15$) to check that we were able to recover the true parameters from a reasonable number of loci and particles per locus (60 loci and 200 particles per locus in our simulations, with various samples). These data are not show here but are available at SAINUDIIN and VÉBER (2017).

Exponential growth model. In Figure 2, we show the likelihood surface for the pair (θ, g), where θ is the scaled per locus mutation rate and g is the population growth rate, based on simulated data from the exponential growth model and $\beta = 0$. The true parameters are $\theta = \phi_1 = 20$, and $g = \phi_2 = 0$ (no growth). Three independent SFS were simulated, corresponding to 3 independent loci and sample size $n = 30$. The $2d$ grid is explored via a quasi Monte Carlo scheme, and for each SFS and each pair (θ, g), the evaluation of the likelihood is based on 1000 particles. The top and bottom left subplot show the likelihood surfaces obtained using only a single locus (labelled 0, 1 and 2), the bottom right subplot shows the product of the likelihoods for all three loci. Of course each SFS corresponds to a single realization of the genealogy with Poissonian mutation of the sample, and so we expect the precision of the likelihood procedure to increase with the number of independent loci considered. This is indeed the case, as shown in Table 1, in which we also make the parameter β vary. There the likelihood estimates are based on 100 particles, which in general should be considered as the minimal number of particles that should be sampled per parameter per locus to ensure a reasonable precision of the estimation via the law of large numbers. However, increasing the number of loci or the number of particles (F, M, T) sampled to

compute the likelihood corresponding to a given locus has a computational cost, and it is therefore important to assess the capabilities of our procedure with reasonable numbers of loci and particles per locus.

[Figure 2 about here.]

[Table 1 about here.]

Table 1 suggests that the number of loci need not be very large for the parameter estimates to be reliable. Table 2 shows that it is also the case when the growth rate is non-zero. Furthermore, small sample sizes are sufficient to detect population growth in the exponential growth model. Indeed, for moderate to large sample sizes the first coalescence events happen very quickly, and at that time the population size has not sufficiently decreased (backwards in time) to have a strong impact on the total edge length in these epochs, and thus on the distribution of mutations on which our procedure is based. In Table 2, we only consider samples of size 2. The parameter β does not play a role in this case and so we set it to 0 in all likelihood calculations. In this case, 30 loci and 1000 particles per locus per SFS are sufficient to detect a deviation from the hypothesis $g = 0$, and furthermore to pin down the appropriate value of g and θ in the short list provided as an example.

[Table 2 about here.]

As concerns the estimation of the tree shape parameter β, sufficiently large deviations from the Kingman case $\beta = 0$, such as $\beta = -1.9$ or $\beta = 50$ are easily detected by our procedure. The reconstruction of the true parameter value is slightly more delicate as it seems that the law of the topology varies slowly with the parameter β. Of course the larger the sample size, the more splits there are in the tree to estimate the different transition probabilities. However, our simulation studies suggest that a sample size of $n = 10$ is already sufficient to obtain close estimates, as long as the number of particles per locus per parameter value is sufficiently large (500 or 1000, for example). See Table 3 for an example.

For values of β closer to 0 a higher per locus mutation rate, i.e. in practice more SNPs in each locus, is necessary to detect a deviation from the topology of Kingman's coalescent and provide a reliable estimate of the parameter β characterizing the balance of the genealogical trees (data available at SAINUDIIN and VÉBER (2017)).

[Table 3 about here.]

The bottleneck model. We also tested the parameter reconstruction in the bottleneck model presented in Section 3 of the Supplementary Material. The data is not shown but can be found at SAINUDIIN and VÉBER (2017). We simulated data corresponding to a recent mild bottleneck, starting at $a = 0.05$, ending at $b = 0.15$, with $N_0 = 1$ and a reduction in population size of $\varepsilon = 0.01$. To reduce the number of parameters, we assumed that the scaled mutation rate θ and the length of the bottleneck $b - a$ was known. The only parameters to reconstruct were then a and ε.

As in the exponential growth model, small sample sizes ($n = 3$) are sufficient to reconstruct the two parameters as long as the number of loci and the number of particles per locus per set of parameters is sufficiently large (30 or 100 loci, 500 or 1000 particles, for an average number of SNPs per locus of the order of 10). Nonetheless, while the starting time of the bottleneck is always well-reconstructed, the population reduction ε tends to be overestimated in general ($\hat{\varepsilon} = 0.1$), even assuming a larger number of SNPs per locus. Note however that as we increase the average number of SNPs per locus, the likelihood surface becomes more and more peaked. Increasing the sample size to 10 or 20 does not seem to improve the precision of the procedure, probably because the high variability in the topology of larger trees is not compensated by the few additional mutations appearing during the bottleneck and carried by larger numbers of individuals in the sample.

DISCUSSION

The procedure developed in this work exploits the huge reduction in the size of the hidden space of genealogical trees to explore when focusing on the optimal tree resolution that fully characterizes the law of the SFS. Furthermore, our sampler produces only tree topologies which are compatible with the observed SFS. This double optimization enables us to compute approximate likelihoods for the parameters describing the demographic history of the population, as well as a parameter β measuring the typical balance of the genealogy of a sample, at a drastically reduced cost compared to procedures based on the leaf-labelled Kingman coalescent. Because the population demographic and structural parameters are shared across (neutral) loci, the per-locus approximate likelihood functions of several hundreds of independent loci can then be combined to bypass the idiosyncratic genealogical history of a single (or a few) loci. For the same reason, dissonant parameter estimates at some loci may enable us to detect outliers, subject to natural selection for example. At the moment, the inference of the parameter β can mainly serve to detect deviations from the assumptions of panmixia and neutrality at the basis of the Kingman coalescent model. A thorough investigation of the effect of different kinds of population structure on the topology of the genealogical tree of a sample could for instance lead to a new and simplified criterion for model selection.

Generalizations. Our inference procedure is flexible and could be generalized in many ways. For instance, the mutation rates could differ between loci to accommodate a potential inhomogeneity in the locus lengths or in the mutation rates along the genome. In addition, because the sampler constructs the compatible SFS histories in an incremental way, we may stop the construction at a step that uses mutations carried by $j > 1$ individuals, for instance if we are only interested in the not-too-recent history of the population. This incremental construction may also lie at the basis of an adaptive exploration of the space of compatible topologies via more sophisticated sequential particle filtering schemes involving genealogical

and interacting systems (Del Moral 2004).

Filtering out non-recombining loci. Our method requires an important pre-processing step to find loci as contiguous blocks of segregatings sites that are free of intra-locus recombination. One could use for example the simple four-gamete test (Hudson and Kaplan 1985), or more complex methods for detecting blocks of loci that are free of intra-locus recombination (for eg. (Posada 2002)). Such filtered loci can then be summarised into SFS and fed into our pipeline for inference. It is important to note that poorly filtered loci with high levels of recombination will tend to give the topological signal of highly unbalanced tree topologies with a local MLE of β close to -2. This is because the fully unbalanced tree is compatible with an SFS that has a positive count at every frequency, i.e. contains singleton, doubleton, ..., i-ton, ..., and $(n-1)$-ton mutations.

Scalable computing framework. Apache Spark, a unified engine for big data processing (Zaharia *et al.* 2016), and the ADAM module (Massie *et al.* 2013) for population genomics in particular, are ideal frameworks for deploying the algorithms developed here for real-world applications at the genomic scale. Rewriting our sageMath/Python codes (Sainudiin and Véber 2017) in Scala will allow for Spark transformations and actions of our algorithms in conjunction with the ETL methods already available in ADAM. Such an undertaking is beyond the scope and resources of this study and we hope that others may pursue such possibilities.

Adding BIM resolution via gene trees. In this work we restricted ourselves to the information in the SFS. Adding additional information can be done systematically using the *partially ordered graph of coalescent experiments* of Sainudiin *et al.* (2011), say from the full binary incidence matrix of mutational patterns across sites and individual sequences, via the haplotype frequencies, and can significantly improve our estimators (see Palacios *et al.* (2017)).

Ethics statement. This statement does not apply to this manuscript.

Data accessibility. The code developed for this work is given in the Supplementary Material. It is also publicly shared at `https://cocalc.com/projects/ac7f397f-eab9-45fc-9278-f486af09ca55/files/FullLikelihoodInferenceSFS.sagews`

Competing interests. We have no competing interests.

Authors' contributions. R.S. and A.V. designed and coded the procedure, studied its properties and drafted the manuscript. All authors gave final approval for publication.

Funding statement. R.S. and A.V. were supported in part by the chaire Modélisation Mathématique et Biodiversité of Veolia Environnement - École Polytechnique - Museum National d'Histoire Naturelle - Fondation X.

LITERATURE CITED

Aldous, D., 2001 Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statist. Sci. *16*(1)**:** 23–34.

Barton, N., A. Etheridge, J. Kelleher, and A. Véber, 2013 Inference in two dimensions: allele frequencies versus lengths of shared sequence blocks. Theor. Pop. Biol. **87:** 105–119.

Beaumont, M., W. Zhang, and D. Balding, 2002 Approximate Bayesian Computation in population genetics. Genetics *162*(4)**:** 2025–2035.

Boitard, S., W. Rodriguez, F. Jay, S. Mona, and F. Austerlitz, 2016 Inferring population size history from large samples of genome-wide molecular data - An Approximate Bayesian Computation approach. PLoS Genetics *12*(3)**:** e1005877.

Bunnefeld, L., L. Frantz, and K. Lohse, 2015 Inferring Bottlenecks from Genome-Wide Samples of Short Sequence Blocks. Genetics *201*(3)**:** 1157–1169.

De Iorio, M. and R. Griffiths, 2004 Importance sampling on coalescent histories. Adv. Appl. Prob. **36:** 417–433.

Del Moral, P., 2004 *Feynman-Kac formulae : genealogical and interacting particle systems with applications*. Springer, New York.

Fearnhead, P. and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

Gattepaille, L., M. Jakobsson, and M. Blum, 2013 Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. Heredity *110*(5)**:** 409–419.

Gutenkunst, R., R. Hernandez, S. Williamson, and C. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics *5*(10)**:** e1000695.

Harris, K. and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genetics *9*(6)**:** e1003521.

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

Ho, S. and B. Shapiro, 2011 Skyline-plot methods for estimating demographic history from nucleotide sequences. Mol. Ecol. Res. **11:** 423–434.

Hobolth, A., M. Uyenoyama, and C. Wiuf, 2008 Importance sampling for the infinite sites model. Statistical Applications in Genetics and Molecular Biology *7*(1)**:** 32.

Hudson, R. and N. Kaplan, 1985 Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences. Genetics *111*(1)**:** 147–164.

Kamm, J., J. Spence, J. Chan, and Y. Song, 2016 Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. Genetics *203*(3)**:** 1381–1399.

Koskela, J., P. Jenkins, and D. Spano, 2015 Computational inference beyond Kingman's coalescent. J. Appl. Probab. *52*(2)**:** 519–537.

MARJORAM, P. and J. WALL, 2006 Fast "coalescent" simulation. BMC Genetics **7:** 16.

MASSIE, M., F. NOTHAFT, C. HARTL, C. KOZANITIS, A. SCHUMACHER, A. D. JOSEPH, and D. A. PATTERSON, 2013, Dec)ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing. Technical Report UCB/EECS-2013-207, EECS Department, University of California, Berkeley.

MCVEAN, G. and N. CARDIN, 2005 Approximating the coalescent with recombination. Phil. Trans. Royal Soc. B **360:** 1387–1393.

MOOERS, A. and S. HEARD, 1997 Inferring evolutionary process from phylogenetic tree shape. The Quarterly Review of Biology *72*(1)**:** 31–54.

NIELSEN, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

NIELSEN, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. **39:** 197–218.

PALACIOS, J., A. VÉBER, J. WAKELEY, and S. RAMACHANDRAN, 2017 BESTT: Bayesian Estimation by Sampling Tajima's Trees. In preparation.

PALACIOS, J., J. WAKELEY, and S. RAMACHANDRAN, 2015 Bayesian nonparametric inference of population size changes from sequential genealogies. Genetics **201:** 281–304.

PETER, B., D. WEGMANN, and L. EXCOFFIER, 2010 Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Mol. Ecol. **19:** 4648–4660.

POSADA, D., 2002 Evaluation of Methods for Detecting Recombination from DNA Sequences: Empirical Data. Molecular Biology and Evolution *19*(5)**:** 708–717.

PYBUS, O., A. RAMBAUT, and P. HARVEY, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155:** 1429–11437.

Roux, C., M. Pauwels, M.-V. Ruggiero, D. Charlesworth, V. Castric, and X. Vekemans, 2012 Recent and ancient signature of balancing selection around the S-locus in *Arabidopsis halleri* and *A. lyrata*. Mol. Biol. Evol. *30*(2)**:** 435–447.

Sainudiin, R., T. Stadler, and A. Véber, 2015 Finding the best resolution for the Kingman-Tajima coalescent: theory and applications,. J. Math. Biol. **70:** 1207–1247.

Sainudiin, R., K. Thornton, J. Harlow, J. Booth, M. Stillman, R. Yoshida, R. Griffiths, G. McVean, and P. Donnelly, 2011 Experiments with the Site Frequency Spectrum. Bulletin of Mathematical Biology *73*(4)**:** 829–872.

Sainudiin, R. and A. Véber, 2017 https://cocalc.com/projects/ac7f397f-eab9-45fc-9278-f486af09ca55/files/FullLikelihoodInferenceSFS.sagews. Public Sage Repository.

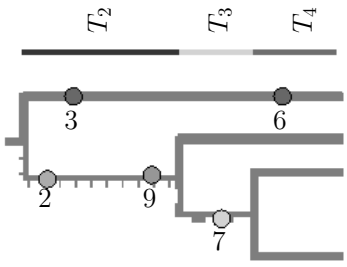
Sawyer, S. and D. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161–1176.

Steinrücken, M., J. Kamm, and Y. Song, 2016 Inference of complex population histories using whole-genome sequences from multiple populations. BioRxiv preprint.

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

Zaharia, M., R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, and I. Stoica, 2016, October)Apache Spark: A Unified Engine for Big Data Processing. Commun. ACM *59*(11)**:** 56–65.

List of Figures



SITES: 1 2 3 4 5 6 7 8 9

IND 1: 0 0 1 0 0 1 0 0 0

IND 2: 0 1 0 0 0 0 0 0 1

IND 3: 0 1 0 0 0 0 1 0 1

IND 4: 0 1 0 0 0 0 1 0 1

Column Sum: 0 3 1 0 0 1 2 0 3

SFS $S = (S_1, S_2, S_3) = (2, 1, 2)$

$$F = \begin{array}{c} \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 3 & 1 & 0 & 1 & 4 \\ 2 & 2 & 1 & 0 & 3 \\ 1 & 4 & 0 & 0 & 2 \end{pmatrix} \end{array}$$

$$M = \begin{array}{c} \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \end{array}$$

Figure 1: The coalescent tree with mutations on the three type of edges subtending 3, 2 and 1 leaves (left), the observed derived mutation incidence matrix with its site frequency spectrum S (middle) and the corresponding SFS history with topology matrix F and mutation matrix M (right). At most one mutation per site under the infinitely-many-sites model are superimposed as a homogeneous Poisson process upon the realization of identical coalescent trees at nine homologous sites labeled $\{1, 2, \ldots, 9\}$ that constitute a non-recombining locus from four individuals labeled $\{1, 2, 3, 4\}$.

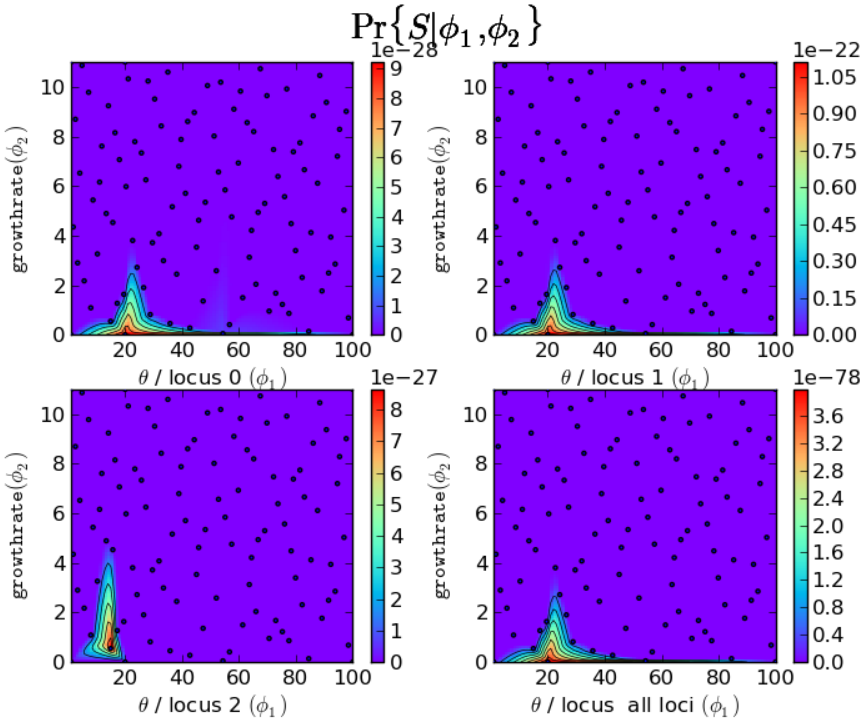


Figure 2: Likelihood surfaces of $\theta = \phi_1$, the per locus scaled mutation rate, and population growth rate $g = \phi_2$ for three loci. True parameters are $\theta = 20$ and $g = 0$, sample size is 30. The black dots shows the points in the parameter space sampled by the quasi-Monte Carlo procedure, except in the region of higher likelihood where the density of sampled parameters is high. The likelihood calculations are based on 1000 particles per parameter per locus.

List of Tables

g	θ	β	Partial log-lkh	Full log-lkh
0	1	-1	-2337.0	-7667.8
0	10	-1	-1576.6	-5226.2
0	100	-1	-3065.1	-10086.1
0	1	0	-2345.4	-7599.9
0	10	0	-1569.6	-5110.1
0	100	0	-3031.3	-10196.8
0	1	10	-2340.0	-7643.8
0	10	10	-1581.8	-5324.1
0	100	10	-3146.0	-10321.1
1	1	-1	-inf	-inf
1	10	-1	-1817.2	-5909.2
1	100	-1	-3047.5	-10086.5
1	1	0	-inf	-inf
1	10	0	-1732.8	-5786.1
1	100	0	-3036.4	-10154.4
1	1	10	-inf	-inf
1	10	10	-1812.2	-5853.3
1	100	10	-3157.5	-10255.5
10	1	-1	-inf	-inf
10	10	-1	-inf	-inf
10	100	-1	-3031.2	-9975.8
10	1	0	-inf	-inf
10	10	0	-inf	-inf
10	100	0	-3019.1	-10014.9
10	1	10	-inf	-inf
10	10	10	-inf	-inf
10	100	10	-3108.1	-10230.8

Table 1: Likelihood of parameters g (growth rate), θ (scaled per-locus mutation rate) and β (balance parameter) under the Beta-splitting model with growth. True parameters are $g = 0$, $\theta = 10$ and $\beta = 0$, the sample size is $n = 15$ and 100 independent SFS have been generated. The mean number of SNPs per locus is 70.05. For each SFS and triplets of parameters, the likelihood estimate is based on 100 particles. The partial log-likelihood is obtained by considering only the first 30 loci, the full log-likelihood uses all 100 SFS. In both cases, the maximum likelihood estimate is obtained for $(\hat{g}, \hat{\theta}, \hat{\beta}) = (0, 10, 0)$

g	θ	Partial log-lkh
0	1	-86.9
0	10	-97.8
0	100	-159.8
1	1	-115.5
1	10	-91.7
1	100	-159.0
10	1	-429.9
10	10	-71.7
10	100	-151.7

Table 2: Likelihood of parameters g and θ under the Beta-splitting model with exponential population growth. Here the true parameter values are $g = 10$, $\theta = 10$ and $\beta = 0$. For a sample of size 2, 100 independent SFS were produced (the mean number of SNPs per locus is 3.82), but only 30 of them are used to compute the approximate likelihoods. 1000 particles are produced per SFS and per pair (g, θ). The most likely parameter values in this short list are $(\hat{g}, \hat{\theta}) = (10, 10)$. The parameter β is not estimated here as a sample of size 2 brings no information on the balance of the genealogy.

g	β	Partial log-lkh
0	-1.9	-489.1
0	0	-651.3
0	10	-746.6
1	-1.9	-478.7
1	0	-656.5
1	10	-750.0
10	-1.9	-525.0
10	0	-731.1
10	10	-826.2

Table 3: Likelihood of parameters g and β under the Beta-splitting model with exponential population growth. Here the true parameter values are $g = 10$, $\theta = 10$ and $\beta = -1.9$. For a sample of size $n = 10$, 100 independent SFS were produced (with a mean number of SNPs per locus of 14.63), but only 30 of them are used to compute the approximate likelihoods. 500 particles are produced per SFS per pair (g, β), using the true value of θ. The procedure detects the imbalance of the genealogical trees and the exponential growth, but underestimates the value of g ($(\hat{g}, \hat{\beta}) = (1, -1.9)$).

Supplementary Material to

Full Likelihood Inference from the Site Frequency Spectrum based on the Optimal Tree Resolution

Raazesh Sainudiin, Amandine Véber

1 Pseudo-code

First, the global process MakeHistory producing a particle is given in Function 1. For the update of the topology, it calls Sstep to insert the edges subtending $n-1$ down to $\lfloor n/2 \rfloor + 1$ leaves, Hstep to insert the edges of size $n/2$ when n is even, Lstep to insert the edges subtending $\lceil n/2 \rceil - 1$ down to 3 leaves, and finally the functions Twostep and Onestep (devoid of randomness) to place the 2- and 1-edges. See the Procedures 4, 5, 6, 7 and 8. All these procedures use the functions IndexSplit, finding the largest block (larger than some quantity given as an input) in the epoch for which it is called, and BetaSplit which computes the required conditioned Beta-splitting probabilities. They are described here in Functions 2 and 3.

Function 1: MakeHistory(A, β, n, S, θ)

Input: A, vector of prior rates for epoch times; β, parameter for Aldous' Beta-splitting model; n, sample size; S, observed SFS; and θ, scaled mutation rate
Output: (F, M, T), an SFS-history of S; and its proposal weights (w, w_1, w_2)

Initialize : $F \leftarrow 0 \in \mathbb{R}^{(n+1)\times(n+1)}$; $M \leftarrow 0 \in \mathbb{R}^{(n+1)\times(n+1)}$; $T \leftarrow 0 \in \mathbb{R}^{n+1}$; $w \leftarrow 1$; $w_1 \leftarrow 1$; $w_2 \leftarrow 1$
foreach $k \in \{2, \dots, n\}$ do T[k] $\leftarrow$ a sample from exponential($A[k]$) random variable;
foreach $k \in \{1, 2, \dots, n-1\}$ do `/* get control sequence C from S */`
  $C[k] \leftarrow 0$; if $S[k] > 0$ then $C[k] \leftarrow 1$;
if $n == 2$ then
  Onestep (β, n, C, F, w); Mutate $(A, n, 1, S[1], \theta, F, M, T, w_1, w_2)$
else if $n == 3$ then
  Twostep (β, n, C, F, w); Mutate $(A, n, 1, S[2], \theta, F, M, T, w_1, w_2)$
  Onestep (β, n, C, F, w); Mutate $(A, n, 2, S[1], \theta, F, M, T, w_1, w_2)$
else
  foreach $j \in \{1, 2, \dots, \lceil n/2 \rceil - 1\}$ do
    Sstep (β, n, j, C, F, w); Mutate $(A, n, j, S[n-j], \theta, F, M, T, w_1, w_2)$
  if (n *is even*) then
    Hstep (β, n, C, F, w); Mutate $(A, n, j, S[n/2], \theta, F, M, T, w_1, w_2)$
  foreach $j \in \{\lfloor n/2 \rfloor + 1, \dots, n-3\}$ do
    Lstep (β, n, j, C, F, w); Mutate $(A, n, j, S[n-j], \theta, F, M, T, w_1, w_2)$
  Twostep (β, n, C, F, w); Mutate $(A, n, n-2, S[2], \theta, F, M, T, w_1, w_2)$
  Onestep (β, n, C, F, w); Mutate $(A, n, n-1, S[1], \theta, F, M, T, w_1, w_2)$
return (F, M, T, w, w_1, w_2);

Function 2: IndexSplit(n, k, V)

Input: n, sample size; k, index; V, vector;
Output: m, largest index greater than or equal to k with $V[k] > 0$, $k-1$ otherwise

$m \leftarrow k - 1$;
foreach $i \in \{n-1, n-2, \dots, k\}$ do
  if $V[i] > 0$ then $m \leftarrow i$; break ;
return m;

Finally, the first half of the procedure Mutate places the S_{n-j} mutations carried by $n-j$ individuals on the newly inserted $(n-j)$-edges, in a multinomial way (if $S_{n-j} > 0$), and updates the importance weight w_1 accordingly. The second half of the procedure samples new Gamma-values for the lengths of the epochs in which there are some $(n-j)$-edges, and updates w_2.

Function 3: BetaSplit(β, m, J)

Input: β, tree shape parameter; m, size of edge split; J, size of largest daughter edge;
Output: I, indicator of whether an m-edge was split into J-edge and $(m-J)$-edge; w, probability of this event

if $J == \lceil m/2 \rceil$ then
  $I \leftarrow 1$; $w \leftarrow 1$;
else
  $p \leftarrow \frac{\lambda_{m,J}}{1 - \sum_{\ell=J+1}^{n-1} \lambda_{m,\ell}}$; $U \sim$ uniform$(0,1)$;
  if $U < p$ then $I \leftarrow 1$; $w \leftarrow p$;
  else $I \leftarrow 0$; $w \leftarrow 1-p$;
return $[I, w]$;

Procedure 4: Sstep(β, n, j, C, F, w)

Data: β, tree shape parameter; n, sample size; j, j-th step; C, control sequence; F, tree topology; and w, weight of F
Result: C, F and w are updated by Sstep

$J \leftarrow n - j$; $F[2, n] \leftarrow n$;
if $\sum_{i=J+1}^{n-1} F[2,i] == 1$ then
  $F[2, J] \leftarrow 0$
else
  if $(C[J] > 0)$ or $((j == \lfloor n/2 \rfloor)$ and $(n \ is \ odd))$ then
    $F[2, J] \leftarrow 1$
  else
    $B \leftarrow$ BetaSplit (β, n, J); $F[2, J] \leftarrow B[0]$; $w \leftarrow w \times B[1]$
if $F[2, J] > 0$ then
  $F[2, 0] \leftarrow J$; $C[J] \leftarrow 0$;
foreach $k \in \{3, j+1\}$ do
  if $\sum_{i=J+1}^{n-1} F[k,i] > 0$ then
    $F[k, J] \leftarrow 0$
  else
    $m \leftarrow$ IndexSplit$(n, J, F[k-1, 0:n])$;
    if $C[J] > 0$ then
      $F[k, J] \leftarrow 1$; $F[k, n] \leftarrow m$; $C[J] \leftarrow 0$;
      if $F[k-1, J] == 0$ then $F[k, 0] = J$;
    else
      if $m < J$ then $F[k, J] \leftarrow 0$;
      else if $m == J$ then
        $U \leftarrow$ sample from uniform$(0,1)$ random variable;
        $q \leftarrow \frac{m-1}{n-k+1}$;
        if $U < q$ then
          $F[k, J] \leftarrow 0$; $F[k, n] \leftarrow J$; $w \leftarrow w \times q$
        else
          $F[k, J] \leftarrow 1$; $w \leftarrow w \times (1-q)$
      else if $J == \lceil m/2 \rceil$ then
        $F[k, J] \leftarrow 1$
      else
        $F[k, n] \leftarrow m$; $B \leftarrow$ BetaSplit (β, m, J); $F[k, J] \leftarrow B[0]$;
        if $F[k, J] == 1$ then $F[k, 0] \leftarrow J$;
        $w \leftarrow w \times B[1]$

Procedure 5: Hstep(β, n, C, F, w)

Data: β, tree shape parameter; n, sample size; C, control sequence; F, tree topology; and w, weight of F
Result: C, F and w are updated by Hstep

$j \leftarrow \lfloor n/2 \rfloor$; $F[2,n] \leftarrow n$;
if $\sum_{i=j+1}^{n-1} F[2,i] == 0$ then $F[2,j] \leftarrow 2$; $F[2,0] \leftarrow j$; $C[j] \leftarrow 0$;
else $F[2,j] \leftarrow 0$;
if $F[2,j] == 0$ then
  foreach $k \in \{3,4,\ldots,j+1\}$ do
    if $\sum_{i=j+1}^{n-1} F[k,i] > 0$ then $F[k,j] \leftarrow 0$;
    else
      $m \leftarrow$ IndexSplit $(n, j, F[k-1, 0:n])$;
      if $C[j] > 0$ then
        $F[k,j] \leftarrow 1$; $F[k,n] \leftarrow m$; $C[j] \leftarrow 0$;
        if $F[k-1,j] == 0$ then $F[k,0] \leftarrow j$;
      else
        if $m < j$ then $F[k,j] \leftarrow 0$;
        else if $m == j$ then
          $U \leftarrow$ sample from uniform$(0,1)$ random variable;
          $q \leftarrow \dfrac{m-1}{n-k+1}$;
          if $U < q$ then $F[k,j] \leftarrow 0$; $F[k,n] \leftarrow j$; $w \leftarrow w \times q$;
          else $F[k,j] \leftarrow 1$; $w \leftarrow w \times (1-q)$;
        else if $j == \lceil m/2 \rceil$ then
          $F[k,j] \leftarrow 1$; $F[k,0] \leftarrow j$;
        else
          $F[k,n] \leftarrow m$; $B \leftarrow$ BetaSplit (β, m, j); $F[k,j] \leftarrow B[0]$;
          if $F[k,j] == 1$ then $F[k,0] \leftarrow j$;
          $w \leftarrow w \times B[1]$;
else
  $F[3,j] \leftarrow 1$; $F[3,n] \leftarrow j$;
  foreach $k \in \{4,5,\ldots,j+1\}$ do
    $m \leftarrow$ IndexSplit $(n, j, F[k-1, 0:n])$;
    if $m < j$ then $F[k,j] \leftarrow 0$;
    else
      $F[k,n] \leftarrow m$; $U \leftarrow$ sample from uniform$(0,1)$ random variable;
      $q \leftarrow \dfrac{m-1}{n-k+1}$;
      if $U < q$ then $F[k,j] \leftarrow 0$; $w \leftarrow w \times q$;
      else $F[k,j] \leftarrow 1$; $w \leftarrow w \times (1-q)$;

Procedure 6: Lstep(β, n, j, C, F, w)

Data: β, tree shape parameter; n, sample size; j, j-th step; C, control sequence; F, tree topology; and w, weight of F

Result: C, F and w are updated by Lstep

$J \leftarrow n - j$; $F[2, n] \leftarrow n$; $F[2, J] \leftarrow F[2, j]$;
if $F[2, J] > 0$ then $C[J] \leftarrow 0$;
foreach $k \in \{3, 4, \ldots, j + 1\}$ do
  $m \leftarrow 0$;
  foreach $i \in \{J + 1, J + 2, \ldots, n - 1\}$ do /* find size of the edge just split, if present */
    if $F[k - 1, i] > F[k, i]$ then $m \leftarrow i$; break ;
  if $m == 0$ then
    if $\left(n - \sum_{i=J}^{n-1}(i \times F[k - 1, i]) - k + 1 + \sum_{i=J}^{n-1} F[k - 1, i]\right) == 0$ then
      $F[k, J] \leftarrow F[k - 1, J] - 1$; $F[k, n] \leftarrow J$
    else if $(C[J] == 0)$ or $\left(\sum_{i=J+1}^{n-1} F[k, i] > 0\right)$ or $(F[k - 1, J] > 1)$ then
      if $F[k - 1, J] == 0$ then $F[k, J] \leftarrow 0$;
      else
        $U \leftarrow$ sample from uniform$(0, 1)$ random variable;
        $q \leftarrow \dfrac{F[k - 1, J] \times (J - 1)}{n - k + 1 - \sum_{l=J+1}^{n-1}(F[k, l] \times (l - 1))}$;
        if $U < q$ then $F[k, J] \leftarrow F[k - 1, J] - 1$; $F[k, n] \leftarrow J$; $w \leftarrow w \times q$;
        else $F[k, J] \leftarrow F[k - 1, J]$; $w \leftarrow w \times (1 - q)$;
    else $F[k, J] \leftarrow F[k - 1, J]$;
  else if $m > 2 \times J$ then
    $F[k, n] \leftarrow m$; $F[k, J] \leftarrow F[k - 1, J] + (F[k, m - J] - F[k - 1, m - J])$;
  else
    $\delta \leftarrow \left(\sum_{i=J+1}^{m-1} F[k, i]\right) - \left(\sum_{i=J+1}^{m-1} F[k - 1, i]\right)$;
    if $(m == 2 \times J)$ and $(\delta == 0)$ then
      $F[k, n] \leftarrow m$; $F[k, J] \leftarrow F[k - 1, J] + 2$; $F[k, 0] \leftarrow J$;
    else if $(m <= 2 \times J)$ and $(\delta > 0)$ then
      $F[k, n] \leftarrow m$; $F[k, J] \leftarrow F[k - 1, J]$;
    else if $(J == \lceil m/2 \rceil)$ and $(\delta == 0)$ then
      $F[k, n] \leftarrow m$; $F[k, J] \leftarrow F[k - 1, J] + 1$; $F[k, 0] \leftarrow J$;
    else if $(m < 2 \times J)$ and $(\delta == 0)$ then
      $F[k, n] \leftarrow m$;
      if $(C[J] > 0)$ and $(F[k - 1, J] == 0)$ then $F[k, J] \leftarrow 1$; $F[k, 0] \leftarrow J$;
      else
        $B \leftarrow$ BetaSplit (β, m, J); $F[k, J] \leftarrow F[k - 1, J] + B[0]$;
        if $B[0] == 1$ then $F[k, 0] = J$;
        $w \leftarrow w \times B[1]$;

Procedure 7: Twostep(β, n, C, F, w)

Data: β, tree shape parameter; n, sample size; C, control sequence; F, topology; and w, weight of F
Result: F and C are updated by Twostep

$j \leftarrow n-2$; $F[2,n] \leftarrow n$; $F[2,2] \leftarrow F[2,j]$;
foreach $k \in \{3,4,\ldots,j+1\}$ do
  $m \leftarrow 0$;
  foreach $i \in \{3,4,\ldots,n-1\}$ do `/* find size of the edge just split, if present */`
    if $F[k-1,i] > F[k,i]$ then $m \leftarrow i$; break;
  if $m == 0$ then
    $F[k,2] \leftarrow F[k-1,2]-1$; $F[k,n] \leftarrow 2$;
  else if $m > 4$ then
    $F[k,n] \leftarrow m$; $F[k,2] \leftarrow F[k-1,2]+F[k,m-2]-F[k-1,m-2]$;
  else if $(m == 4)$ and $(F[k,3]-F[k-1,3] == 0)$ then
    $F[k,n] \leftarrow m$; $F[k,2] \leftarrow F[k-1,2]+2$; $F[k,0] \leftarrow 2$;
  else if $(m == 4)$ and $(F[k,3]-F[k-1,3] > 0)$ then
    $F[k,n] \leftarrow m$; $F[k,2] \leftarrow F[k-1,2]$;
  else if $m == 3$ then
    $F[k,n] \leftarrow m$; $F[k,2] \leftarrow F[k-1,2]+1$; $F[k,0] \leftarrow 2$;
$C[j] \leftarrow 0$;

Procedure 8: Onestep(β, n, C, F, w)

Data: β, tree shape parameter; n, sample size; C, control sequence; F, topology; and w, weight of F
Result: F and C are updated by Onestep

$j \leftarrow n-1$; $F[2,n] \leftarrow n$; $F[2,1] \leftarrow F[2,j]$;
foreach $k \in \{3,4,\ldots,j\}$ do
  $m \leftarrow 0$;
  foreach $i \in \{3,4,\ldots,n-1\}$ do
    if $F[k-1,i] > F[k,i]$ then $m \leftarrow i$; break;
  if $m > 2$ then
    $F[k,1] \leftarrow F[k-1,1]+(F[k,m-1]-F[k-1,m-1])$; $F[k,n] \leftarrow m$;
  else
    $F[k,1] \leftarrow F[k-1,1]+2$; $F[k,0] \leftarrow 1$; $F[k,n] \leftarrow 2$;
  $F[n,1] \leftarrow n$;
$F[n,n] \leftarrow 2$; $F[n,0] \leftarrow 1$; $C[j] \leftarrow 0$;

Procedure 9: Mutate$(A, n, j, s, \theta, F, M, T, w_1, w_2)$

Data: A, rates of *a priori* exponential epoch times; n, sample size; j, j-th step; s, mutations carried by $n-j$ individuals; θ, scaled mutation rate; F, topology; M, mutation matrix; T, epoch times; w_1, weight of M and w_2, weight of T

Result: M, T, w_1 and w_2 are updated by Mutate

$J \leftarrow n - j$;

foreach $i \in \{0, 1, \ldots, n\}$ do $M[i, J] \leftarrow 0$;

if $s \neq 0$ then

$$M[2 : j+1, J] \sim \text{multinomial}\left(s, \left(\frac{F[2,J] \times T_2}{\sum_{\ell=2}^{j+1}(F[\ell,J] \times T_\ell)}, \frac{F[3,J] \times T_3}{\sum_{\ell=2}^{j+1}(F[\ell,J] \times T_\ell)}, \ldots, \frac{F[j+1,J] \times T_{j+1}}{\sum_{\ell=2}^{j+1}(F[\ell,J] \times T_\ell)}\right)\right);$$

$$w_1 \leftarrow w_1 \times s! \prod_{i=2}^{j+1} \frac{1}{M[i,J]!} \left(\frac{F[i,J] \times T_i}{\sum_{\ell=2}^{j+1}(F[\ell,J] \times T_\ell)}\right)^{M[i,J]};$$

foreach $k \in \{2, 3, \ldots, j+1\}$ do

if $F[k, J] > 0$ then

$$a \leftarrow 1 + \sum_{i=J}^{n-1} M[k, i]; \quad b \leftarrow A[k] + \theta \sum_{i=J}^{n-1} F[k, i];$$

$T[k] \sim \text{gamma}(a, b)$;

$$w2 \leftarrow w2 \times \frac{b^a}{\Gamma(a)} T[k]^{a-1} \exp(-bT[k])$$

2 Exposition of the Algorithm when $n = 8$

Let us detail how MakeHistory (the full procedure constructing a tree topology, a mutation matrix and an epoch time vector compatible with a given SFS) works.

Suppose $n = 8$ and the observed SFS is $S = (5, 2, 0, 0, 1, 0, 2)$. Let us see how our sampler constructs a tree with mutations based on this information. We assume that $\beta = 0$ to simplify the expression of the probabilities related to the topology of the tree. The control sequence created at the beginning of the procedure tells us which edge sizes need to be seen in the tree. Here, it is thus equal to $C = (1, 1, 0, 0, 1, 0, 1)$.

Recall that during step j, the edges subtending $n - j$ leaves are placed in the tree.

2.1 Topology Matrix F.

We start from an $(n+1) \times (n+1)$ matrix whose entries are all equal to 0 (indexed from 0 to n), and a proposal weight $w = 1$.

j = 1: Since $C(7) = 1$, Sstep forces the presence of a 7-edge in the only epoch at which such an edge is possible, that is epoch 2. Hence, $F(2, 7) := 1$ and since a 7-edge now exists in the tree, $C(7) := 0$. The largest edge created during the first split has size 7 and the edge split during this step subtended 8 leaves by construction, and so $F(2, 0) := 7$ and $F(2, 8) := 8$. On the other hand, we do not know yet the size of the largest edge created by the split of the 7-edge, so that $F(3, 8) := 7$ but $F(3, 0)$ remains equal to 0 for now. This call of Sstep ends here.

j = 2: Because of the presence of a 7-edge at epoch 2 (i.e., $F(2, 7) > 0$), there cannot be an edge of size 6 at this epoch and $F(2, 6) = 0$. Next, $C(6) = 0$ and so the algorithm may or may not split the 7-edge into a 6- and a 1-edge. Let us say that it creates no 6-edges, which happens with probability 1/3 when $\beta = 0$. Hence, $F(3, 6) = 0$ and $F(3, 0)$ (the size

of the largest edge created by the split of the 7-edge) remains equal to 0 too. Also, the weight w associated to the tree is multiplied by the above probability, that is $w := 1/3$ after this step. This call of Sstep ends here.

j = 3: Again, there can be no 5-edge at epoch 2. Next, $C(5) = 1$ and so the 7-edge needs to be split into a 5-edge and a 2-edge. Since at this step Sstep updates only the entries corresponding to the 5-edges, we obtain $F(3,5) := 1$, $F(3,0) := 5$ and $C(5) := 0$. In epoch 4, IndexSplit (see Section A) gives the size of the largest edge present in epoch 3, that is 5. Since a 5-edge has already been placed in the previous epoch, the presence or absence of this 5-edge in epoch 4 is random. With probability 4/5, we decide that it is absent, and so $F(4,5) = 0$ and $w := 1/3 \times 4/5 = 4/15$. This means that the 5-edge at epoch 3 was split and so $F(4,8) := 5$. This call of Sstep stops here.

j = n/2 = 4: Because $\sum_{l=5}^{7} F(2,l) > 0$ and $\sum_{l=5}^{7} F(3,l) > 0$, there cannot be a 4-edge in epochs 2 and 3. Next IndexSplit returns 5, the size of the largest edge present in epoch 3. Since $\sum_{l=5}^{7} F(4,l) = 0$ and $C(4) = 0$, the presence of a 4-edge in epoch 4 is random. Let us say that such an edge is created by the split of the 5-edge, which happens with probability 1/2. Thus, $F(4,4) := 1$, $F(4,0) := 4$ and $w := 4/15 \times 1/2 = 2/15$. Using the same procedure (with IndexSplit returning 4 now), the 4-edge is not split at the beginning of the next epoch with probability 1/4, so that $F(5,4) := 1$ and $w := 1/30$. Finally, since there cannot be a 4-edge in epoch $k \geq 6$, Hstep forces the split of this edge, $F(6,4) = 0$ and $F(6,8) := 4$ (while w remains the same). This call of Hstep stops here.

j = 5: First, $F(2,3) := F(2,5) = 0$. Next, in each epoch, Lstep looks for the size m of the edge split just before this epoch, if it has been already decided. This size is $m = 7$ for epoch 3. Since $7 > 2 \times 3$, the largest edge created by this split has already been decided and $F(3,3) := F(2,3) + F(3,7-3) - F(2,7-3) = 0$. In epoch 4, $m = 5$ and $F(4,4) - F(3,4) > 0$ (a 4-edge has been created), and so $F(4,3) := F(3,3) = 0$. Then, $m < 4$ and $F(4,3) = 0$, hence $F(5,3)$ remains equal to 0 with probability 1. The edge split at the beginning of epoch 6 has size $m = 4$, we do not know yet the size of the largest edge created by this split and $C(3) = 0$, hence the presence of a 3-edge in epoch 6 is random. Let us say that it is absent, which happens with probability 1/3: we thus have $F(6,3) = 0$ and $w := 1/90$. This call of Lstep stops here.

The last two steps (placing 2- and 1-edges) are fully deterministic and so the final weight of the tree topology obtained is $w = 1/90$.

j = 6: First, $F(2,2) := F(2,6) = 0$. Next, in each epoch, Twostep again looks for the size m of the edge split at its beginning, if such an edge already exists (otherwise $m = 0$). Hence, in epoch 3 we have $m = 7 > 4$ and so $F(3,2) := F(2,2) + F(3,5) - F(2,5) = 1$. In epoch 4, $m = 5$ and $F(4,2) := F(3,2) + F(4,3) - F(3,3) = 1$. In epoch 5, $m = 0$ and so a 2-edge needs to be split: $F(5,2) = 0$ and $F(5,8) = 2$. In epoch 6, $m = 4$ and since no 3-edge was created by this split, we have $F(6,2) := F(5,2) + 2 = 2$ and $F(6,0) = 2$. In epoch 7, $m = 0$ and so $F(7,2) = F(6,2) - 1 = 1$ and $F(7,8) = 2$. Finally, $F(8,2)$ remains equal to 0 (there are only 1-edges) and $F(8,8) := 2$.

j = 7: Onestep considers each split, epoch by epoch, and checks whether the number of 1-edges remains the same, or increases by 1 or 2 (the latter being the consequence of the split of a 2-edge). Hence, $F(2,1) = F(3,1) := 1$, $F(4,1) := 2$, $F(5,1) = F(6,1) := 4$, $F(7,1) := 6$ and $F(8,1) := 8$. Also, $F(5,0) = F(7,0) = F(8,0) := 1$.

The tree topology we obtain is thus (recall that $F(k,8)$ gives the size of the edge split at the beginning of epoch k and $F(k,0)$ that of the largest edge created by this split):

$$\begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{ccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{pmatrix} 7 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 8 \\ 5 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 7 \\ 4 & 2 & 1 & 0 & 1 & 0 & 0 & 0 & 5 \\ 1 & 4 & 0 & 0 & 1 & 0 & 0 & 0 & 2 \\ 2 & 4 & 2 & 0 & 0 & 0 & 0 & 0 & 4 \\ 1 & 6 & 1 & 0 & 0 & 0 & 0 & 0 & 2 \\ 1 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix} \end{array}$$

2.2 Mutation Matrix M and Epoch Time Vector T.

We present the construction of the mutation matrix M and of the epoch time vector T in a separate paragraph for the sake of clarity, but in fact the mutations carried by $n-j$ individuals in the sample are placed just after the j-th partial update of the topology matrix F.

We start from an $(n+1)\times(n+1)$ matrix M whose entries are all 0, and an $(n+1)$ vector T such that T_k is a realization of an exponential random variable with parameter A_k.

For j ranging from 1 to n, after the j-th update of the topology we first check whether there are $(n-j)$-mutations to place (i.e., $S(n-j)>0$). If it is the case, we use the distribution of the $(n-j)$-edges just obtained and the current value of the epoch time vector T to give a weight W to each epoch and distribute the mutations in a multinomial way. For example, for $j=1$, there is only one edge in epoch 2 subtending $n-1=7$ leaves, and this edge is split at the beginning of epoch 3. Consequently, the only possible allocation of the $S_7=2$ mutations carried by 7 individuals is to declare that $M(2,7)=2$ and $M(k,7)=0$ for $k>2$. The time $T(2)$ is then updated by taking an independent sample from a $\mathcal{G}(1+2, A_2+\theta)$ distribution. The importance weight w_2 is multiplied by the likelihood of the sampled value, and no other epoch times are updated.

Likewise, the only edge subtending 5 leaves is placed during step $j=3$ in epoch 3 and it is split at the beginning of epoch 4. This imposes that $M(3,5)=1$ and $M(k,5)=0$ for $k\neq 3$, and $T(3)$ is replaced by an independent sample from a $\mathcal{G}(1+1, A_3+\theta)$ distribution. The weight w_2 is updated accordingly.

As concerns the $S_2=2$ mutations carried by 2 individuals, during the $(n-2)nd$ update of the topology we inserted 1 2-edge in epoch 3, 1 in epoch 4 (the continuation of that in epoch 3), 2 in epoch 6, 1 in epoch 7 and 0 in epochs 2, 5 and 8. This gives the weights

$$W(3)=T(3),\ W(4)=T(4),\ W(6)=2T(6),\ W(7)=T(7)$$

to the epochs in which we see some 2-edges, and $W(k)=0$ otherwise. Writing $L_2=\sum_{k=2}^{n} W(k)$ for the total length of 2-edges in the partial topology constructed up to step $n-2$ (included), we then sample the second column of the mutation matrix according to the following multinomial distribution:

$$\big(M(2,2), M(3,2), \ldots, M(8,2)\big) \sim \text{Multinomial}\left(S_2\,;\ \frac{W(2)}{L_2}, \frac{W(3)}{L_2}, \ldots, \frac{W(8)}{L_2}\right).$$

We multiply the importance weight w_1 by the probability of the mutation allocation sampled. We then update the values of $T(k)$ by taking independent samples from $\mathcal{G}\big(1+\sum_{l=2}^{7} M(k,l), A_k+\theta\sum_{l=2}^{7} M(k,l)\big)$ distributions, only for $k=3,4,6,7$. The weight w_2 is multiplied by the likelihood of this 4-sample of times.

We proceed in the same way to arrange the 5 mutations carried by a single individual on the final topology and update the epoch times and importance weights accordingly. In the end, a possible mutation matrix created by the procedure is the following:

$$\begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

3 Likelihoods of Simple Parametric Models

The simulations presented in the article are based on the Kingman coalescent topology, $\beta = 0$. The corresponding probability of a topology is given in (**??**). We thus only need to specify the likelihood of the vector of epoch times resulting from the sampling of a particle. We provide three classical examples.

3.1 Homogeneous coalescence rates.

We start with the case where the coalescence rates are homogeneous, i.e., constant over every epoch. For every $k \in \{2,\dots,n\}$, $\mathcal{A}_k$ gives the parameter of the exponential random variable corresponding to epoch k, and so

$$\mathbb{P}^{(\mathcal{A})}(T) = \prod_{k=2}^{n} \big(\mathcal{A}_k e^{-\mathcal{A}_k T_k}\big). \tag{1}$$

As a major example, choosing $\mathcal{A}_k = k(k-1)/(2N_0)$ for a given $N_0 > 0$ corresponds to assuming that the epoch times are those of the Kingman coalescent with effective population size N_0.

3.2 Inhomogeneous coalescence rates.

To ease the notation, let us first define the cumulated epoch times $\overline{T}_k$ by $\overline{T}_{n+1} = 0$ and for every $k \in \{2,\dots,n\}$,

$$\overline{T}_k = T_n + T_{n-1} + \cdots + T_k.$$

In words, $\overline{T}_k$ is the total amount of time during which the sample has at least k ancestors. Suppose the coalescence rate of k lineages at time t is given by a function $\mathcal{A}_k(t)$. Under

this general assumption, the probability of a given vector of epoch times is

$$\mathbb{P}^{(\mathcal{A})}(T) = \prod_{k=2}^{n} \mathcal{A}_k(\overline{T}_k) e^{-\int_{\overline{T}_{k+1}}^{\overline{T}_k} \mathcal{A}_k(s)ds}. \tag{2}$$

Next we consider two special cases of inhomogeneous coalescence rates, assuming that each pair of blocks coalesces independently at the same instantaneous rate (as in the Kingman coalescent). They correspond to a population which is growing exponentially according to a growth rate parameter g, and to a population which has undergone a bottleneck that can be described with four parameters.

3.2.1 Exponential Growth.

The probability density of the epoch times under an exponential population growth with parameter g (forwards in time) is obtained from (2) and

$$\mathcal{A}_k(t) = \frac{k(k-1)}{2e^{-gt}},$$

as

$$\begin{aligned}\mathbb{P}^{(g)}(T) &= \prod_{k=2}^{n} \left(\frac{k(k-1)}{2} e^{g\overline{T}_k} e^{-\int_{\overline{T}_{k+1}}^{\overline{T}_k} \frac{k(k-1)}{2} e^{gs} ds} \right) \\ &= \prod_{k=2}^{n} \left\{ \frac{k(k-1)}{2} \exp\left(g\overline{T}_k - \frac{k(k-1)}{2g} e^{g\overline{T}_{k+1}} \left(e^{gT_k} - 1 \right) \right) \right\}.\end{aligned}$$

3.2.2 Bottleneck.

Suppose the population size is N_0, but shrunk to εN_0 (with $\varepsilon > 0$ not necessarily less than 1 in the following calculation) between $a > 0$ and $b > a$ units of times in the past. See Fig. 1 for an illustration.

To find the likelihood of the vector of epoch times $T = (T_2, \ldots, T_n)$ in this scenario, let us further define the indices at which the transition between the different phases occurs:

$$\begin{aligned} k_* &:= \min\{k \in \{2, \ldots, n+1\} \, : \, \overline{T}_k < a\} \\ k_{**} &:= \min\{k \in \{2, \ldots, n+1\} \, : \, \overline{T}_k < b\}. \end{aligned}$$

For ease of computation, let us split this scenario into 3 situations.

First, if $T_n > b$ then all the coalescence events occurred before the bottleneck, and the only epoch time whose likelihood is modified by the bottleneck is T_n. Hence, using the general formula obtained in the previous section we have

$$\mathbb{P}^{(N_0,\varepsilon,a,b)}(T) = \frac{1}{N_0}\binom{n}{2} \exp\left\{ -\frac{1}{N_0}\binom{n}{2}\left(a + \frac{b-a}{\varepsilon} + T_n - b \right) \right\} \times \prod_{k=2}^{n-1} \left[\frac{1}{N_0}\binom{k}{2} e^{-\frac{1}{N_0}\binom{k}{2}T_k} \right].$$

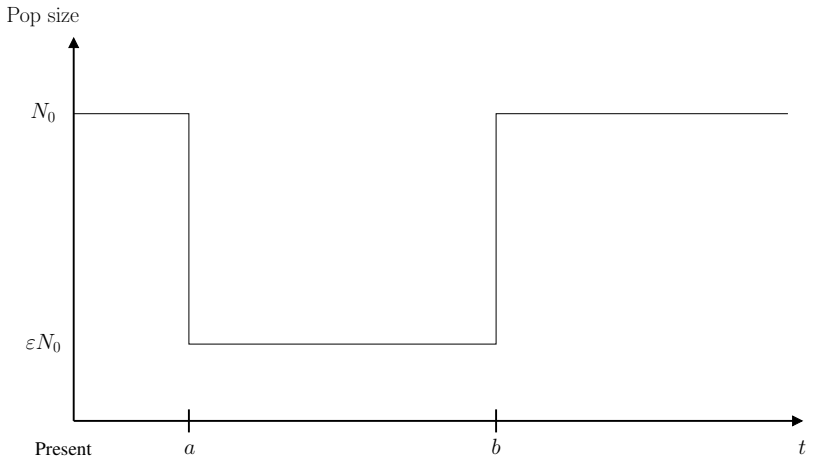


Figure 1: Bottleneck model with four parameters, N_0 effective population size at present, ε factor of reduction of population size, $a < b$ start and end time of the bottleneck period (time running backwards).

Second, if $T_n \leq b$ but $k_* = k_{**}$ (meaning that no epoch ends between a and b in the past), the only epoch time whose likelihood is modified by the bottleneck is that during which $k_* - 1$ lineages are ancestral to our sample. This gives

$$\mathbb{P}^{(N_0,\varepsilon,a,b)}(T) = \prod_{k=k_*}^{n} \left[\frac{1}{N_0}\binom{k}{2} e^{-\frac{1}{N_0}\binom{k}{2}T_k}\right] \times \prod_{k=2}^{k_*-2} \left[\frac{1}{N_0}\binom{k}{2} e^{-\frac{1}{N_0}\binom{k}{2}T_k}\right] \times \frac{1}{N_0}\binom{k_*-1}{2} \exp\left\{-\frac{1}{N_0}\binom{k_*-1}{2}\left(a - \overline{T}_{k_*} + \frac{b-a}{\varepsilon} + \overline{T}_{k_*-1} - b\right)\right\}.$$

To simplify this expression a bit, we may notice that $a - \overline{T}_{k_*} + \frac{b-a}{\varepsilon} + \overline{T}_{k_*-1} - b = (b-a)(1/\varepsilon - 1) + T_{k_*-1}$.

Finally, in none of the above cases we obtain that

$$\mathbb{P}^{(N_0,\varepsilon,a,b)}(T) = \prod_{k=k_*}^{n} \left[\frac{1}{N_0}\binom{k}{2} e^{-\frac{1}{N_0}\binom{k}{2}T_k}\right] \times \prod_{k=k_{**}}^{k_*-2} \left[\frac{1}{\varepsilon N_0}\binom{k}{2} e^{-\frac{1}{\varepsilon N_0}\binom{k}{2}T_k}\right] \times \prod_{k=2}^{k_{**}-2} \left[\frac{1}{N_0}\binom{k}{2} e^{-\frac{1}{N_0}\binom{k}{2}T_k}\right] \times \frac{1}{\varepsilon N_0}\binom{k_*-1}{2} \exp\left\{-\frac{1}{N_0}\binom{k_*-1}{2}\left(a - \overline{T}_{k_*} + \frac{\overline{T}_{k_*-1} - a}{\varepsilon}\right)\right\} \times \frac{1}{N_0}\binom{k_{**}-1}{2} \exp\left\{-\frac{1}{N_0}\binom{k_{**}-1}{2}\left(\frac{b - \overline{T}_{k_{**}}}{\varepsilon} + \overline{T}_{k_{**}-1} - b\right)\right\}.$$