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

Raazesh Sainudiin, Amandine Véber

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

Raazesh Sainudiin[1], Amandine Véber [2]

June 19, 2018

[1]Department of Mathematics, Uppsala University, Uppsala, Sweden
[2]CMAP, CNRS, École Polytechnique, Palaiseau, France

Running Head: Likelihood of Non-recombining Site Frequency Spectrum

Keywords: 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, 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 always compatible with the observed SFS.

To produce the particles, 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 estimate the likelihood of a population scenario based on a large collection of independent SFS by an importance sampling scheme, using the (unconditional) distribution of the genealogies under this scenario when the latter is available. When it is not available, we instead compute the joint likelihood of the tree balance parameter β assuming that the tree topology follows Aldous' Beta-splitting model, and of the demographic scenario determining the distribution of the inter-coalescence times or epoch times in the genealogy of a sample, in order to at least distinguish different equivalence classes of population scenarios leading to different tree balances and epoch times. Simulation studies are conducted to demonstrate the capabilities of the approach with publicly available code.

1 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 of 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, 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 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 [GJB13] 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 introduce a novel importance sampler, which 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. $\mathbb{P}(S \,|\, F, M, T) > 0$. Here, the epoch time T_k is the amount of time during which the sample has exactly k ancestors. When the distribution (not conditioned on the observed SFS) of the hidden genealogy of a sample under the family of population scenarios of interest is known, we can use these triplets to estimate the corresponding likelihood function *given the observed SFS*. In particular, it allows us to compute the

likelihood of a given scenario in a (parametric or not) family of stochastic past population size trajectories $(N_t,\, t \geq 0)$, assuming that the genealogy of a sample is given by the Kingman coalescent with fluctuating population size, by specializing to the case $\beta = 0$ in the Beta-splitting model lying at the heart of our method (cf. Section 2.1). See Section 3 for some parametric examples, including exponential growth and bottlenecks. If the unconditional law of the sample genealogy is not known, we can instead resort to supposing that the tree topology is well-described by Aldous' Beta splitting model (see below), which specifies a measure on tree balance through a single parameter β, and then estimate the joint likelihood of the parameter β and of a demographic model dictating the law of the epoch times. Under a given model of demography specifying the distribution of epoch times, our procedure would then allow one to distinguish between different equivalence classes of historical population scenarios up to their tree topology distributions specified by different values of β. Although we illustrate how to apply such a *Beta-splitting demographic model* in a semi-parametric spirit upon data simulated from a complicated scenario using `ms` [Hud02] in Section 3.5, we emphasize that it is not the point of this paper to study the effect of any given structural scenario on the β-specific shapes of the genealogical trees, or to inquire which scenarios can be distinguished based on their balance parameters.

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 [SSV15]. 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 [SSV15], each binary tree shape represents 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 separately consider SFS-compatible genealogical histories at each one of a number of independent non-recombining loci — a natural framework for subsequent locus-specific outlier (non-neutrality) detection.

Second, the sampler needs to be informed of population-level processes equally affecting the genealogies at all independent loci only through (i) their tree topology or shape via the parameter $\beta \in (-2, \infty)$ in Aldous' Beta-splitting model [Ald01], introduced below to account for topologically sensitive processes including certain types of population structure and sampling schemes, and (ii) their branch lengths via a vector of *a priori* mean epoch times which typically summarize the demographic scenario experienced by the population. 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 non-neutral 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 what follows), 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, but none of them are able to compute the full likelihood of a given SFS at a non-recombining locus. The *Poisson Random Field* approach [SH92, Nie00, GHWB09] 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 [Nie00], while the sampling probability is obtained by solving a Wright-Fisher diffusion with selection as in [GHWB09]. Our procedure generalizes this approach. Indeed, the Poisson random field methodology imposes that each locus should be a single segregating site that is at an infinite recombinational distance from all other segregating sites, and thereby misses the shared genealogical signal at linked sites. In contrast, the likelihood that we compute does not assume that all SNPs are independent at the locus of interest, nor does it assume that the probability of a given mutation being carried by exactly k individuals is the *average* (over many tree realizations) of the proportion of the tree length corresponding to the total length of k-edges. Instead, our procedure can handle SFS data with fully-linked SNPs within each given locus, by constructing locus-specific particles whose genealogical and mutational histories are compatible with the SFS at the locus. We can then extend it to several independent loci which are free of intra-locus recombination to reduce the variance of the likelihood-based estimator. Now, 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 [Nie00], 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 site.

In [BFL15], 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 [HS11]. 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., [PRH00]). However, this will typically not be the case when the individual mutation rate is low and the SFS contains only a few mutations. [HD08] 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, [HN13] 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 [MC05] and the related SMC′ model of [MW06]. 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 [BEKV13] 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 [SKS16] 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). [PWR15] 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* [Taj83, SSV15] considerably enhances the exploration of this component of the state space during the MCMC step, a point which is pushed further by [PVWR18]. 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 [BZB02] is now routinely used in a wealth of studies. Recently [BRJ$^+$16] 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 [PWE10] 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. As described below, this controlled Markov process is the step-by-step construction of a tree topology and a vector of epoch times, controlled by a vector of presence or absence of mutations carried by each number of individuals. This vector is used to ensure that at every step $k \geq 1$, (at least) one edge subtending $n-k$ leaves is necessarily placed in the tree whenever the component S_{n-k} of the SFS is positive. The resulting tree is thus always compatible with the given SFS.

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 [Nie05, RPR$^+$12]. However, when one wants to infer the locus-specific history provided by the SFS, to our knowledge there are no known methods analogous to the importance samplers available for inference from the full *binary incidence matrix* or BIM data [FD01, DIG04, HUW08, KJS15, KSCS16], except the naive importance sampler via controlled Markov chains in [STH$^+$11]. 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 parameters in (i) demographic models with

analytical expressions for likelihood, such as, exponential growth or bottleneck models (under Kingman coalescent with $\beta = 0$) and in (ii) Beta-splitting demographic models with $\beta \in (-2, \infty)$ under a simple demographic scenario specified by exponential rates for epoch times.

Such Beta-splitting demographic models are meant to be a simple semi-parametric approximation of an equivalence class of much more complex historical processes with no analytically tractable likelihood function, such as time-varying migration structures, admixture times and/or sub-population specific demographies. Such an approximation may be an acceptable alternative to simulation-intensive inference methods such as ABC that require one to envision specific high-dimensional parametric families of historical scenarios to simulate from. 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 paper is organized as follows. In Section 2, we progressively present our sampler by introducing the *a priori* laws used for the tree topology (§2.1) and the vector of epoch times (§2.2), then by precisely defining the triplet (F, M, T) describing a genealogical tree with mutations at the optimal resolution (§2.3), and finally by giving a full description of how the sampler produces such a 'particle' (§2.4). In Section 3, we perform some simulation studies for a few examples of Beta-splitting demographic models (Kingman's coalescent with exponential growth in §3.2 and with bottlenecks in §3.3), and we estimate β for a complex population history simulated using `ms` in §3.5. Some generalizations of this approach to the inference of demographic and structural scenarios are discussed in Section 4. The proof-of-concept code for the sampler and the likelihood procedure as well as its integration with `ms` is publicly available at [SV18]. The detailed pseudo-code can be found in Section 1 of the Supplementary Material.

2 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 $k \in \{2, \dots, n\}$ 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,

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

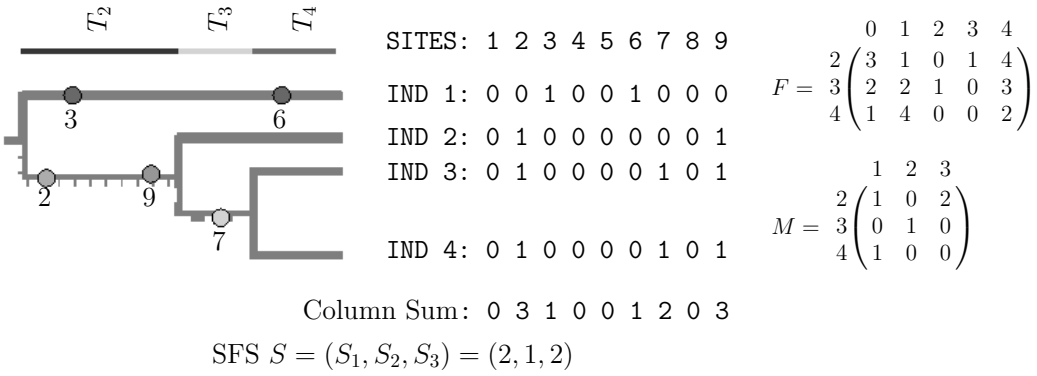

Figure 1: The coalescent tree with mutations on the three type of edges subtending 3, 2 and 1 leaves (left), the observed derived mutation incidence matrix with its site frequency spectrum S (middle) and the corresponding SFS history with topology matrix F and mutation matrix M (right). At most one mutation 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 $\{1, 2, 3, 4\}$.

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 [STH$^+$11], 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 [Ald01], our simple one-parameter model of various phenomena affecting tree shape (see Section 2.1).

Similarly,

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

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 [STH$^+$11], 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.

2.1 *A Priori* Laws for the Topology

Our sampler uses a modification of the Beta-splitting model of [Ald01]. Indeed, the Aldous Beta-splitting model is a one-parameter family of random cladograms (i.e., binary splitting trees in which the order of the splits is not specified) which has the advantage of containing several of the most classical null models for tree shapes used in phylogeny reconstruction ([MH97]), 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, with

$$\lambda_{b,x} = \frac{2 - \mathbf{1}_{\{x=b/2\}}}{b-1}. \tag{9}$$

Choosing β close to -2 gives rise to comb-like trees, while $\beta \gg 1$ produces highly balanced tree topologies [Ald01]. Thus the Beta-splitting model gives us a one-parameter family spanning the whole range of possible tree balances. In our modified version, we use the same expression for the probability of each split, but we also specify the order of the splits by enforcing a stochastic rule derived from the random split order in the Kingman coalescent.

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' *stick-breaking* construction ([Ald01]); 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

$$\mathbb{P}_\beta(F) = \prod_{k=2}^{n} \left(\frac{F(k-1,m_k)(m_k-1)}{n-k+1} \lambda_{m_k,\ell_k} \right). \tag{10}$$

In particular, when $\beta = 0$, using (9) we see that if $\neg(F)$ denotes the number of splits such that $\ell_k \neq m_k/2$, we indeed recover the probability

$$\mathbb{P}_0(F) = \frac{2^{\neg(F)}}{(n-1)!} \prod_{k=2}^{n} F(k-1,m_k) \tag{11}$$

of the unvintaged and sized Kingman coalescent (see Proposition 11 in [SSV15]).

2.2 *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

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

Other examples of parametric models are given in Section 3.1.2. 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 standard 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

$$\frac{\lambda^k}{\Gamma(k)} t^{k-1} e^{-\lambda t} \mathbf{1}_{\{t>0\}} \tag{12}$$

(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 $(k+m, \lambda+\theta)$. (More precisely, if T follows a $\mathcal{G}(k, \lambda)$ distribution and if, conditionally on the value of T, M is a Poisson random variable with parameter θT, then it follows from Bayes' rule that

$$T \,|\, \{M = m\} \sim \mathcal{G}(k+m, \lambda+\theta).$$

See Section 3 in the Supplementary Material for a short proof.)

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: If the current *a priori* law on some epoch time T is $\mathcal{G}(k, \lambda)$ and our algorithm places m new mutations in this epoch, the *a posteriori* distribution on T (which will become the *a priori* distribution in the next step of the procedure) is $\mathcal{G}(k+m, \lambda+\theta)$).

2.3 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

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

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 (an expression for the density of $\mathbb{P}_{A,\beta,\theta}$ is given in §3.1.3). 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

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

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) has density

$$q_{A,\beta,\theta}[(F, M, T) \mid S] \propto w w_1 w_2, \tag{13}$$

where w, w_1 and w_2 are the weights output by the sampler corresponding to F, M and T, respectively. A *particle* refers to such an S-compatible SFS history and the corresponding 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.

2.4 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 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 (an example with $n = 5$ is presented below):

(i) **Updating the topology:** We update the topology matrix F and the 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 (since the presence or absence of m-edges in every epoch is already fully determined for every $m > n-j$). At the moment when this m-edge splits, we also know that $n-j$ is the biggest size possible for the largest 'daughter' edge (otherwise the creation of an edge subtending more than $n-j$ leaves at the time of this split 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 weight w by the corresponding conditional probability. If the $(n-j)$-edge is indeed created, we now have to decide in which epoch it is split and gives rise to two smaller edges. To this end, let us observe that all the other edges present at the same time in the tree are necessarily smaller (recall that $n-j > n/2$). Consequently, 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. Thus, if the state of the epoch time vector just before this step is $(T_2^{(n-j+1)}, \ldots, T_n^{(n-j+1)})$, the total length of $(n-j)$-edges in any epoch k in the current state of the tree is $F(k, n-j)T_k^{(n-j+1)}$, and for each of the S_{n-j} mutations to be placed (independently) on the tree, the probability that it occurs in epoch k is

$$\frac{F(k, n-j)T_k^{(n-j+1)}}{\sum_{\kappa=2}^{n} F(\kappa, n-j)T_\kappa^{(n-j+1)}}.$$

We also update the current value of the 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 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.

2.5 Example

Take $n = 5$ and suppose that the observed SFS is $S = (4, 0, 1, 0)$. We give an example of the construction of a particle with $\beta = 0$ and $A_k = k(k-1)/2$, i.e., the parameters corresponding to the Kingman coalescent. The topology and mutation matrices F and M are initialized at 0, and the epoch time vector T is initialized at a vector of samples from independent random variables with distribution $\text{Exp}(k(k-1)/2)$. The weights w, w_1, w_2 are initialized at 1. Finally, the control vector is set to be $(1, 0, 1, 0)$ since there are mutations carried by 1 and 3 individuals, but no mutations are carried by 2 or 4 individuals in the sample. Recall that the columns 0 and n of F record respectively the size of the largest daughter edge and the size of the edge split at the beginning of the k-th epoch.

j = 1 (mutations carried by $n - 1 = 4$ individuals):
Updating F: Since $S_4 = 0$, there are no constraints on the presence of a 4-edge in the tree. Using (9), with probability 1/2 we choose to create a 4-edge in epoch 2, thus setting $F(2, 4) = 1$. We also multiply w by 1/2. Since between epochs 1 and 2 we have split the ancestral 5-edge and produced a (largest) daughter edge subtending 4 leaves, we also set $F(2, 5) = 5$ and $F(2, 0) = 4$. Epoch 2 being the only epoch in which the presence of a 4-edge is possible, the update of F based on S_4 stops here and $F(k, 4)$ remains equal to 0 for all $k > 2$.

Updating M: Since $S_4 = 0$, there are no mutations to place on the tree and w_1 is not updated.
Updating T: Because the (only) 4-edge placed in epoch 2 of the tree carries no mutation, we update T_2 only, by replacing its current value by a sample from a Gamma$(1, 1+\theta)$ variable (i.e., the law of an Exp(1)-random variable T_2 conditioned on Poisson$(\theta T_2) = 0$). We also multiply w_2 by the density of the Gamma distribution at the sampled point.

j = 2 (mutations carried by $n - 2 = 3$ individuals):
Updating F: In epoch 2, we have already placed a 4-edge and so a 3-edge is impossible and $F(2, 3)$ remains equal to 0. Next, in epoch 3, since $C_3 = 1$ the split of the 4-edge in the previous epoch needs to lead to the creation of a 3-edge (which otherwise could not appear later in the tree), and so we set $F(3, 3) = 1$ with probability 1 (and we set $C_3 = 0$ since an edge able to accommodate the 3-mutation has been included in the tree, and so the presence of another 3-edge is not compulsory). We also set $F(3, 5) = 4$ and $F(3, 0) = 3$ to record the fact that between epochs 2 and 3, we have split a 4-edge and created a largest daughter edge of size 3. Since a 3-edge cannot appear in a later epoch, the update of F stops here and $F(k, 3)$ remains equal to 0 for $k = 4, 5$.
Updating M: The only epoch containing a 3-edge is epoch 3, and so the mutation carried by 3 individuals needs to appear in this epoch. We thus set $M(3, 3) = S_3(= 1)$ and keep the other $M(k, 3)$ equal to 0. The weight w_1 is not updated since the chosen mutation placement has probability 1.
Updating T: Since epoch 3 is the only epoch containing a 3-edge, T_3 is the only time about which we have some new information for a possible update. As $M(3, 3) = 1$, we update T_3 by sampling a new time from a Gamma$(2, 3+\theta)$ distribution and we multiply w_2 by the density of the Gamma distribution at the sampled point.

j = 3 (mutations carried by $n - 3 = 2$ individuals):
Updating F: In epoch 2 (resp., 3), the presence of the 4- (resp., 3-)edge prevents the presence of a 2-edge, so that $F(2, 2) = 0 = F(3, 2)$. The 3-edge being the one split between epochs 3 and 4 (since $F(4, 3) - F(3, 3) = -1$), it necessarily leads to the creation of a 2-edge in epoch 4. Hence, we set $F(4, 2) = 1$ and we do not need to update w since this move has probability 1. We also set $F(4, 5) = 3$ and $F(4, 0) = 2$. The 2-edge is the only possible edge to split in the next (and last) epoch, and so $F(5, 2) = 0$ (and w does not need to be updated).
Updating M: $S_2 = 0$, and so there are no mutations carried by two individuals to place in the tree. Consequently, neither M nor w_1 is updated.
Updating T: The only 2-edge in the tree is that appearing in epoch 4, on which $S_2 = 0$

mutations are placed, and so the current value of T_4 is replaced by a sample from a Gamma$(1, 6+\theta)$ random variable and w_2 is multiplied by the value of the density of this Gamma distribution at the sampled point.

j = 4 (mutations carried by $n-4=1$ individuals):
Updating F: As for the 2-edges, there is no randomness in the placement of 1-edges in the topology. In epoch 2, the creation of the 4-edge from the split of the ancestral 5-edge imposes that $F(2,1)=1$. This edge cannot be split further, and so its remains in every later epoch. In epoch 3, the creation of the 3-edge from the split of a 4-edge leads to the creation of a new 1-edge, so that $F(3,1)=2$. Likewise, $F(4,1)=3$ and since the 2-edge in epoch 4 is split into two 1-edges, we necessarily have $F(5,1)=5$, $F(5,5)=2$ and $F(5,0)=1$. The weight w is not updated since all these changes have probability 1.
Updating M: For each of the $S_1=4$ mutations carried by one individual, independently of each other, the probability p_k that it is placed in epoch k is given by

$$p_2=\frac{T_2}{T},\ p_3=\frac{2T_3}{T},\ p_4=\frac{3T_4}{T},\ p_5=\frac{5T_5}{T},$$

with $T=T_2+2T_3+3T_4+5T_5$. We thus draw the column $(M(k,1))_{2\le k\le 5}$ from a Multinomial$(4;p_2,p_3,p_4,p_5)$ distribution, and we multiply the weight w_2 by the probability of this draw. For instance, $M(2,1)=0$, $M(3,1)=1$, $M(4,1)=1$ and $M(5,1)=2$.
Updating T: We have placed no mutations carried by one individual on the single 1-edge in epoch 2, and so T_2 is replaced by a sample from a Gamma$(1,1+2\theta)$ r.v. and w_2 is multiplied by the value of the density of the Gamma distribution at the sampled point. Likewise, T_3 is replaced by an independent sample from a Gamma$(3,3+3\theta)$ variable, T_4 is replaced by a sample from a Gamma$(2,6+4\theta)$ variable and T_5 is resampled from a Gamma$(3,10+5\theta)$ variable. The weight w_2 is also multiplied by the values of densities of these Gamma distributions at the sampled points.
The sampler then outputs the topology and mutation matrices

$$(F_{k,j})_{\substack{2\le k\le 5\\0\le j\le 5}}=\begin{pmatrix}4&1&0&0&1&5\\3&2&0&1&0&4\\2&3&1&0&0&3\\1&5&0&0&0&2\end{pmatrix}\quad,\quad (M_{k,j})_{\substack{2\le k\le 5\\1\le j\le 4}}=\begin{pmatrix}0&0&0&0\\1&0&1&0\\1&0&0&0\\2&0&0&0\end{pmatrix},$$

which correspond to the totally unbalanced topology with possible arrangement of mutations displayed in Figure 2, as well as the final state of the epoch time vector T and of the weights w ($=1/2$ in this simple example), w_1 and w_2.

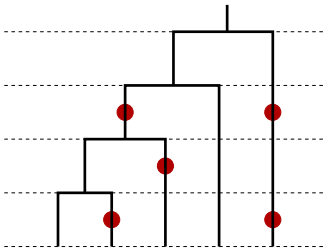

Figure 2: Tree with mutations corresponding to the result of the sampling described in the example. The F-matrix fully characterizes the tree topology, while the mutation pattern shown here is only one instance of the possible mutation placements corresponding to the matrix M. For instance, there is only one 3-edge in epoch 3 on which to place the mutation corresponding to $(M(3,3)=1)$, but the mutation corresponding to $M(3,1)=1$ may be carried by any of the two singleton edges in epoch 3.

3 Simulations and likelihood computations

In this section, we first describe our standard importance sampling procedure to compute likelihood functions in two cases. In §3.1.1, we focus on the case where we have an explicit expression for the law of the genealogy of a sample (*not* conditioned on the observed SFS) under the family of scenarios considered. Some examples are given in §3.1.2. In §3.1.3, we suggest that when no such expression is available, trying to find the most likely demographic Beta-splitting model can at least enable us to distinguish between demographic and structural scenarios leading to different tree balances and epoch time distributions. We then explore the properties of our sampler through simulations from the Kingman coalescent with two standard models of demography (for which we have explicit formulae for the law of the genealogy), the exponential growth model in §3.2, and the bottleneck model in §3.3. 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). To test our approach in the case where the law of the genealogies is not explicitly known, in §3.5 we use `ms` (integrated to our code) to produce a series of SFS under a complex demographic

scenario of multiple populations exchanging migrants and we used the methodology of §3.1.3 to explore the tree balance and branch length distribution under this scenario.

These results and more of them are available (and can be reproduced) at [SV18].

3.1 Likelihood computations

To estimate the likelihood of a given demographic and structural scenario, we use the standard importance sampling methodology [TK10]. It may be improved in different ways, in particular the step-by-step construction of each 'particle' may be the basis of an adaptive importance sampling scheme with resampling and partial 'mutation' of the particles (see the last sections in [TK10]). We adopt two different approaches, depending on whether the unconditional distribution of genealogical trees is available or not for the class of scenarios considered. In both cases, to produce the genealogies with mutations we need to inform the sampler with a balance parameter β, a mutation rate θ and a vector of rates for the *a priori* exponential distributions of the epoch times.

The mutation rate and balance parameters, if unknown, may be considered as parameters to be estimated by the procedure. As concerns the vector A giving the rates in the *a priori* distribution of the epoch time vector, if no explicit expressions exist for the class of scenarios considered, the rates can be obtained by a round of simulations of genealogical trees under the scenario considered without conditioning on the observed SFS. In this way, we can estimate the mean epoch times under this scenario and take the inverse of these averages as A_k.

3.1.1 Scenarios for which a likelihood function is available

Suppose that we consider a class of demographic and structural scenarios $\mathcal{F}$ under which the law $\mathbf{P}_f$ of the genealogy of a sample (not conditioned on the observed SFS S) is known. In this case, we can evaluate the likelihood of a scenario f by sampling m independent particles $[(F_i, M_i, T_i), (w^{(i)}, w_1^{(i)}, w_2^{(i)})]$ and by using the fact that for m large enough

$$\mathbf{P}_f(S) \approx \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\{SFS(M_i)=S\}} \frac{\mathbf{P}_f((F_i, M_i, T_i))}{w^{(i)} w_1^{(i)} w_2^{(i)}} = \frac{1}{m}\sum_{i=1}^{m} \frac{\mathbf{P}_f((F_i, M_i, T_i))}{w^{(i)} w_1^{(i)} w_2^{(i)}}, \tag{14}$$

where we abuse notation and write also $\mathbf{P}_f$ for the density function of the distribution $\mathbf{P}_f$ (observe that the law of the epoch time vector is absolutely continuous with respect to Lebesgue measure). The last equality comes from the fact that the sampler produces particles which are always compatible with the SFS given as an argument. Assuming the

infinitely many site model with Poissonian mutation occurring at rate θ, we can further decompose the expressions in the numerators as

$$\mathbf{P}_f((F_i, M_i, T_i)) = \mathbf{P}_f((F_i, T_i))\mathbf{P}_f(M_i|(F_i, T_i)).$$

We know the (density of the) law of the genealogical tree (F, T) under $\mathbf{P}_f$, and furthermore the law of M knowing (F, T) is the product over all epochs k and numbers of carriers j of the probability of that a Poisson$(\theta F(k, j)T_k)$ variable is equal to $M(k, j)$:

$$\begin{aligned}\mathbf{P}_f(M|(F, T)) &= \prod_{k=2}^{n}\prod_{j=1}^{n-1} e^{-\theta F(k,j)T_k}\frac{(\theta F(k, j)T_k)^{M(k,j)}}{M(k, j)!} \\ &= e^{-\theta L}\theta^{\sum_i S_i}\prod_{k=2}^{n}\prod_{j=1}^{n-1}\frac{(F(k, j)T_k)^{M(k,j)}}{M(k, j)!}, \end{aligned} \tag{15}$$

where $L = \sum_{k,j} F(k, j)T_k$ is the total length of the tree. Consequently, every term in (14) can be computed to obtain an approximation to the likelihood of $f \in \mathcal{F}$. Note that $\mathcal{F}$ does not need to be a parametric family here.

In our simulation studies, we use a version of (14) in which $\mathbf{P}_f((F_i, M_i, T_i))$ is replaced by $\mathbf{P}_f((F_i, S, T_i))$. This modification was motivated by the fact that for large sample sizes, the product in (15) may have very small values compared to the probability of the SFS knowing the tree, because of the large number of possible mutation placements corresponding to the same SFS. Although we do not have a precise justification for it, the modified likelihood procedure enables us to recover the true parameters among a grid of parameter values (which we did not take to be too large in the examples presented below for want of resources to develop an optimised version of our code).

3.1.2 Likelihoods of simple parametric models

To compute the numerator of the importance weight in the likelihood estimation procedure described in §3.1.1, we first need the law of the topology of the genealogy. In the simplest models of population genetics, this topology is that of the Kingman coalescent and its distribution is recalled in (11). This value is then multiplied by the likelihood of the vector of epoch times resulting from the sampling of a particle in the class of demographic models considered. In this section, we provide three classical examples for the computation of these general epoch time distributions.

In all that follows, we write $\mathcal{A}_k(t)$ for the instantaneous rate of pairwise coalescence at time t in the past when there are k extant lineages. This should not be confused with

the parameter A_k in the *a priori* law of the epoch time vector of our sampler, which is supposed to be the inverse of the average time during which the sample has k ancestors in the demographic model considered (of course both quantities are related).

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

$$\mathbb{P}_{\mathcal{A}}(T) = \prod_{k=2}^{n} \left(\mathcal{A}_k e^{-\mathcal{A}_k T_k}\right). \tag{16}$$

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.

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 instantaneous coalescence rate of k lineages at time t is given by a function $\mathcal{A}_k(t)$. Under this general assumption, the probability density of a given vector of epoch times is

$$\mathbb{P}_{\mathcal{A}}(T) = \prod_{k=2}^{n} \mathcal{A}_k(\overline{T}_k) e^{-\int_{\overline{T}_{k+1}}^{\overline{T}_k} \mathcal{A}_k(s)ds}. \tag{17}$$

Next we consider two (parametric) 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.

Exponential Growth.
The probability density of the epoch times under an exponential population growth with parameter g (forwards in time) is obtained from (17) and

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

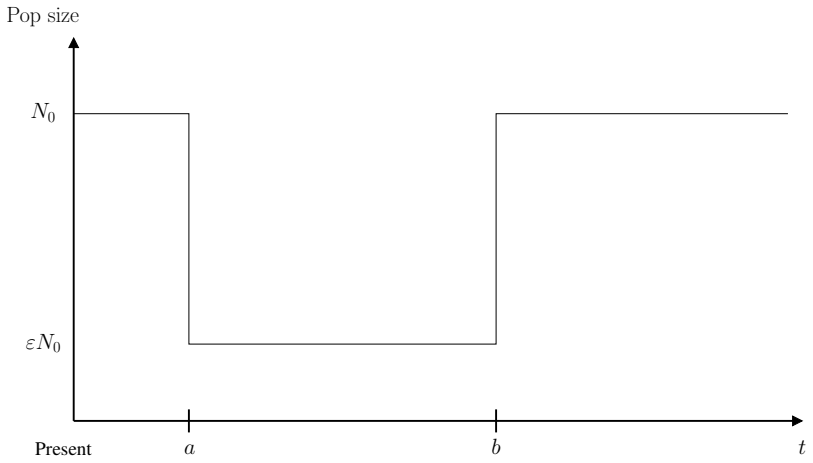

Figure 3: Bottleneck model with four parameters, N_0 effective population size at present, ε factor of reduction of population size, $a < b$ start and end time of the bottleneck period (time running backwards).

as

$$\begin{aligned}\mathbb{P}_g(T) &= \prod_{k=2}^{n} \left(\frac{k(k-1)}{2}\, e^{g\overline{T}_k} e^{-\int_{\overline{T}_{k+1}}^{\overline{T}_k} \frac{k(k-1)}{2} e^{gs} ds} \right) \\ &= \prod_{k=2}^{n} \left\{ \frac{k(k-1)}{2}\, \exp\left(g\overline{T}_k - \frac{k(k-1)}{2g} e^{g\overline{T}_{k+1}} \left(e^{gT_k} - 1 \right) \right) \right\}. \end{aligned} \tag{18}$$

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. 3 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:

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

That is, $k_* - 1$ is the number of extant lineages at the moment when the bottleneck starts (backwards in time), and $k_{**} - 1$ is the number of extant lineages when it ends. 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

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

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

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

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

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

Indeed, by definition of t_* and t_{**}, the beginning (backwards in time) a of the bottleneck happens during epoch $k_* - 1$, and the end happens during epoch $k_{**} - 1$. Hence, during epochs 2 to $k_{**} - 2$ the population size is constant equal to N_0, and likewise during epochs k_* to n. During epochs k_{**} to $k_* - 2$, it is constant equal to εN_0. The last two terms in the product correspond to the probability density of the epoch times T_{k_*-1} and $T_{k_{**}-1}$, during which the population size changes from N_0 to εN_0 and from εN_0 back to N_0.

3.1.3 Scenarios for which a likelihood function is not available

Suppose that we do not have an analytic expression for the (unconditional) distribution of the genealogies under the scenarios of interest. In this case, we can resort to finding the Beta-splitting demographic model with exponential epoch times which best fits the data. Recall our notation $\mathbb{P}_{A,\beta,\theta}$ for the probability measure under which (i) the genealogy of a sample follows the incremental Beta-splitting model with parameter β defined in §2.1, (ii) with independent exponentially distributed epoch times whose parameters (or rates) are given by the vector A, and (iii) with mutations occurring on the tree at rate θ. That is, using (10), the expression for the density of exponential random variables and (15), the density of a given particle (F, M, T) under $\mathbb{P}_{A,\beta,\theta}$ (which we also write $\mathbb{P}_{A,\beta,\theta}$ for simplicity) is given by

$$\begin{aligned}\mathbb{P}_{A,\beta,\theta}((F,M,T)) = &\prod_{k=2}^{n}\left(\frac{m_k-1}{n-k+1}\lambda_{m_k,\ell_k}\right)\prod_{k=2}^{n}\left(A_k e^{-A_k T_k}\right)\\ &\times e^{-\theta L}\theta^{\sum_i S_i}\prod_{k=2}^{n}\prod_{j=1}^{n-1}\frac{(F(k,j)T_k)^{M(k,j)}}{M(k,j)!},\end{aligned} \tag{19}$$

where L is the total length of the tree.

Using the same method as in the previous paragraph, we can estimate the likelihood of (A, β, θ) as

$$\mathbb{P}_{A,\beta,\theta}(S) \approx \frac{1}{m}\sum_{i=1}^{m}\frac{\mathbb{P}_{A,\beta,\theta}((F_i,M_i,T_i))}{w^{(i)}w_1^{(i)}w_2^{(i)}}.$$

Note that if the law of the genealogy with mutations output by the sampler is close to the distribution $\mathbb{P}_{A,\beta,\theta}(\cdot \mid S)$, then we do recover that the importance weights satisfy

$$\frac{\mathbb{P}_{A,\beta,\theta}((F_i,M_i,T_i) \mid S)}{w^{(i)}w_1^{(i)}w_2^{(i)}} \approx 1.$$

In this case, each ratio $\mathbb{P}_{A,\beta,\theta}((F_i,M_i,T_i))/(w^{(i)}w_1^{(i)}w_2^{(i)})$ is very close to $\mathbb{P}_{A,\beta,\theta}(S)$ and only a small number of particles is sufficient to obtain a precise approximation for $\mathbb{P}_{A,\beta,\theta}(S)$. However, because the construction of a particle (and its weights) is incremental in the number j of mutation carriers while the computation of the probability $\mathbb{P}_{A,\beta,\theta}((F,M,T))$ is incremental in the epoch labels k, we do not have an analytic way to check how close the sampler's distribution is to the conditional law $\mathbb{P}_{A,\beta,\theta}(\cdot \mid S)$.

3.2 Exponential growth model

To check the performances of our procedure, we first considered the one-parameter model of the Kingman coalescent with exponential growth (with growth rate $g \geq 0$). The data was obtained by simulating a series of pairs of tree topology and epoch time vector using the law (11) of the Kingman coalescent for the topology, and a vector of epoch times with density function (18). These simulations were carried out using standard procedures that have been integrated into our code, see [SV18]. For each genealogy, an independent SFS was produced using a specified scaled per locus mutation rate θ. The series of SFS was then used as a basis for the importance sampling experiments, using the approach detailed in §3.1.1.

In Figure 4, we show the likelihood surface for the pair (θ, g). The true parameters are $\theta = \phi_1 = 20$, $g = \phi_2 = 0$ (no growth). Three independent SFS were simulated, corresponding to 3 independent loci and sample size $n = 30$. The 2-dimensional grid over θ and g is explored with β fixed at 0 (i.e., using the true *a priori* distribution on topologies) 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 two and bottom-left subplots 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 maximum likelihood estimator to improve with the number of independent loci considered. This is indeed the case in another simulation study involving 100 loci over a coarse proof-of-concept grid, as shown in Table 1, in which we also make the parameter β vary, assuming that the law of the topology belongs to the larger family of Beta-splitting models and therefore replacing (11) by (10) in the likelihood computations. (This can be seen either as defining a (β, g, θ)-parametric model and computing the likelihood function for this family of models in the spirit of §3.1.1, or as fitting our data to the best Beta-splitting model with exponential growth as a demographic scenario in the spirit of §3.1.3.)

Here the likelihood estimates are based only 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. Table 1 suggests that the number of loci need not be very

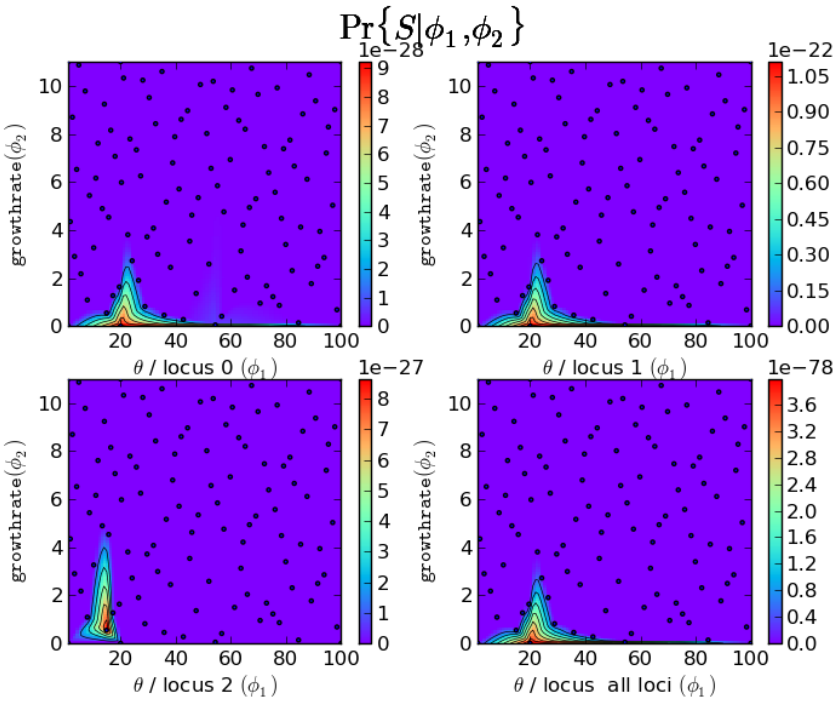

Figure 4: Likelihood surfaces of $\theta = \phi_1$, the per locus scaled mutation rate, and population growth rate $g = \phi_2$ for three loci. True parameters are $\theta = 20$, $g = 0$ and $\beta = 0$ with a sample size of 30. The black dots shows the points in the parameter space sampled by the quasi-Monte 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.

large for the parameter estimates to be reliable.

Table 2 shows another simulation study 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 to correctly estimate the true value of g and θ in the short list provided as an example.

In order to assess the reliability of our maximum likelihood inference procedure, we simulated a set of 100 independent SFS under the standard Kingman coalescent sce-

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	10	-1	-1817.2	-5909.2
1	100	-1	-3047.5	-10086.5
1	10	0	-1732.8	-5786.1
1	100	0	-3036.4	-10154.4
1	10	10	-1812.2	-5853.3
1	100	10	-3157.5	-10255.5
10	100	-1	-3031.2	-9975.8
10	100	0	-3019.1	-10014.9
10	100	10	-3108.1	-10230.8

Table 1: Log-likelihood (log-lkh) 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 $(\hat{g}, \hat{\theta}, \hat{\beta}) = (0, 10, 0)$. The parameter points in the Cartesian grid for $(g, \theta, \beta) \in \{0, 1, 10\} \times \{1, 10, 100\} \times \{-1, 0, 10\}$ with likelihood estimates of $-\infty$ in double precision are not shown.

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: Log-likelihood (log-lkh) of parameters g and θ under the Kingman coalescent topology with exponential population growth. Here the true parameter values are $g = 10$, and $\theta = 10$. 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 log-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 in the particle sampling was always set to 0, since a sample of size 2 brings no information on the balance of the genealogy.

nario ($\beta = 0 = g$, $\theta = 10$) with sample size $n = 4$. We then produced 10 estimates for the likelihood of the true parameter, and for that of a close but erroneous parameter $(\beta, \theta, g) = (0, 10, 1)$, using different numbers of particles per locus in the likelihood estimation. As shown in Table 3, as the number of particles per locus increases, the distribution of the log-likelihood estimate for each parameter becomes more and more concentrated. Furthermore, the log-likelihood estimates of the true parameter $(0, 10, 0)$ become significantly larger than those of the erroneous parameter $(0, 10, 1)$. Due to the order of magnitude difference in log-likelihood values between distant parameter values, we found that even 20 particles are often sufficient to quickly explore the relative log-likelihoods over a coarse grid of parameters. A more careful exploration of the parameter space will require a larger number of particles per locus (about 100 particles at least) in general.

Regarding 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. Milder deviations of β from 0 such as -1 or 10 can be detected with 30 loci with sample size of 15 as shown in Table 4. Distinguishing smaller deviations of β from 0 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

	Nb particles	1	10	100	500	1000
$(0, 10, 0)$	mean log-lkh	-1126.7	-649.2	-516.8	-454.5	-425.1
	std. err.	13.3	10.5	4.0	3.3	2.6
$(0, 10, 1)$	mean log-lkh	-1599.5	-903.9	-690.7	-613.8	-583.7
	std. err.	35.3	10.6	5.9	4.6	3.4

Table 3: Concentration of the log-likelihood estimates as the number of particles per locus used in the procedure increases. We simulated 100 independent SFS under the standard Kingman coalescent scenario (true parameter $(\beta, \theta, g) = (0, 10, 0)$). Based on this series of SFS, we produced 10 estimates of the log-likelihood of the true parameter and of some close but erroneous parameter $(0, 10, 1)$, using different numbers of particles per locus (per parameter) in the importance sampling procedure. For each parameter value and each number of particles per locus, we report the mean and standard error of the 10 replicates.

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 (from 100 to 1000, for example).

Finally, we checked that by considering larger mutation rates, we can reconstruct the population growth rate more finely. In Table 5, we took $g = 2$, $n = 15$, $\theta = 100$ and $\beta = 0$ (Kingman coalescent), which gave an (unrealistic) mean number of segregating sites of 364.35 over the 60 independent SFS which were simulated. Assuming that we know θ, our procedure is able to recover the true value of g among a rather fine grid, and to check that the underlying trees had a balance corresponding to the Kingman coalescent.

3.3 The bottleneck model

We also investigated parameter estimation in the bottleneck model presented in §3.1.2, using again the methodology of §3.1.1. The results are not shown but can be found at [SV18]. We simulated data corresponding to the Kingman coalescent going through 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 with 500 or 1000 particles per locus, for an average number of SNPs per locus of the order of 10). Nonetheless, while

β	θ	g	log-lkh true $\beta = 10$	log-lkh true $\beta = -1$
-1	1	0	-2426.0	-2485.4
-1	10	0	-1807.3	**-1595.6**
-1	100	0	-3353.2	-3009.7
0	1	0	-2386.0	-2497.5
0	10	0	-1764.6	-1629.1
0	100	0	-3296.4	-3054.6
10	1	0	-2318.4	-2524.5
10	10	0	**-1651.9**	-1721.7
10	100	0	-3295.3	-3233.8
-1	100	10	-3334.1	-3029.8
0	100	10	-3336.5	-3063.6
10	100	10	-3275.0	-3196.2

Table 4: Log-likelihood (log-lkh) of parameters β, θ and g under the Beta-splitting model with exponential population growth. Here the true parameter values are $g = 0$, $\theta = 10$ with (i) an unbalanced $\beta = -1$ and (ii) a more balanced $\beta = 10$. SFS at 30 independent loci were simulated for a sample size of 15. The log-lkh was only based on 20 particles per SFS per triplet $(\beta, \theta, g) \in \{-1, 0, 10\} \times \{1, 10, 100\} \times \{0, 10\}$. The procedure detects the true parameters corresponding to the maximum log-lkh values in bold. Rows with log likelihood estimates of $-\infty$ in double precision are not shown.

β	g	Log-lkh
-1	0	-4908
0	0	-4770
10	0	-4734
-1	1	-4859
0	1	-4758
10	1	-4755
-1	2	-4938
0	2	**-4693**
10	2	-4719
-1	3	-5014
0	3	-4846
10	3	-4837
-1	5	-5950
0	5	-5768
10	5	-5729
-1	10	-13897
0	10	-13682
10	10	-∞

Table 5: Log-likelihood (log-lkh) of parameters g (growth rate), and β (balance parameter) under the Beta-splitting model with growth. True parameters are $g = 2$, $\theta = 100$ and $\beta = 0$, the sample size is $n = 15$ and 60 independent SFS have been generated. The mean number of SNPs per locus is 364.35. For each SFS and pair of parameters, the likelihood estimate is based on 200 particles. The maximum likelihood estimate is obtained for $(\hat{\beta}, \hat{g}) = (0, 2)$.

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. 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.

3.4 Integration with `ms` and systematic parameter search

We integrated Hudson's `ms` program [Hud02] with a minor modification to output SFS in order to validate against our simulators and estimators. For example, we were able to estimate the true parameters (same parameter setting as in Table 1) when the SFS were directly simulated from `ms` as opposed to our simulators. This integration with `ms` will allow one to simulate SFS data under rather complex demographic and structural scenarios and simply fit this data to a Beta-splitting demographic model as done in §3.5.

We also confirmed that the true parameters are recovered through a stochastic global optimization algorithm that evolves a population of parameters towards the MLE that concentrates about the true parameters [SP97], albeit at a larger computational cost. The stochastic optimizer used here is based on a set of points in the parameter space that evolve towards the MLE. When this population concentrates about the MLE (so that the mean of the set of points in the parameter space is close to the MLE and the variance is very small) then the algorithm is said to have converged to the true parameter, i.e. the set of parameters that the data were simulated from. We were able to recover the MLE for several settings for the beta-splitting exponential growth model. These results are not given here as they are fully reproducible from [SV18]. Having demonstrated that systematic parameter searches can be done (especially with optimized version of our code and/or with more computational resources), we henceforth resort to coarse grids of parameter values to quickly illustrate our likelihood computations with our proof-of-concept code under different simulation scenarios.

3.5 Fitting to the simplest Beta-splitting demographic model

Consider a rather complex historical scenario of a stepping-stone model with a recent barrier (as given in Fig. 3 of the documentation for `ms` [Hud02], see Section 4 in the Supplementary Material for details). There are six subpopulations that exchange migrants in a stepping-stone model. At a time $T = 2$ time units in the past a barrier to gene flow

β	θ	Log-lkh
-1	10	-12062
-1	100	-6101
-1	1000	-11882
0	10	-11864
0	100	-5796
0	1000	-11594
5	10	-11737
5	100	**-5735**
5	1000	-11524
10	10	-11734
10	100	-5812
10	1000	-11546

Table 6: Log-likelihood (log-lkh) of parameters β (balance parameter) and θ (scaled per-locus mutation rate) under the Beta-splitting model with constant population size 1 (*i.e.*, with a pairwise coalescence rate equal to 1). See the text for more details. The maximum likelihood estimate is obtained for $(\hat{\beta}, \hat{\theta}) = (5, 100)$. The pairs of parameters $(-1, 1), (0, 1), (5, 1), (10, 1)$ all had likelihood estimates of $-\infty$ in double precision.

arose, such that no further gene flow occurs between subpopulation 3 and subpopulation 4.

No analytical likelihood expression is available for this complex historical scenario. We can now fit the 100 SFS loci with sample size $n = 15$ and per locus mutation rate $\theta = 10$ to the simplest Beta-splitting model with constant population size. The log-likelihood is estimated over a coarse grid of parameters over β and θ using 1000 particles per locus. As shown in Table 6, the maximum log-likelihood value corresponded to more balanced topologies with $\hat{\beta} = 5$ due to the sampling scheme and the population structure and larger value of θ due to longer time to coalescence caused by the barrier with $\hat{\theta} = 100$. Note that when β is set to 0 the most likely θ is proportional to the product of the mutation rate per locus and the effective population size N_e under the standard Kingman coalescent. Our simplest Beta-splitting model considered here adds an additional tree-balance parameter to the classical setting with constant population size (for details see [SV18]) and is over 60 log-likelihood units better than the model with $\beta = 0$ and $\theta = 100$. We purposely keep the demography of the fitted β-splitting model as simple as possible here to illustrate that some population histories involving complex population structures can be explained significantly by a single tree-balance parameter.

4 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 (even with our non-optimized 'proof of concept' code). 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.

The next step is to develop this procedure into an optimized code, enabling us to perform a thorough quantitative comparison with the existing methods for reconstructing demographic parameters. However, we stress again that one of the novelties of this approach is that it is able to compute (approximate) full likelihoods for SFS at non-recombining loci, which none of the existing methods are able to do (except [BFL15] for very small sample sizes).

4.1 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 com-

patible topologies via more sophisticated sequential particle filtering schemes involving genealogical and interacting systems [DM04]. To make the semi-parametric family of Beta-splitting models richer and fit more complex tree topological distributions, we can allow β's to change at each epoch or be drawn from a distribution with minimal change to the sampler, provided we have a much larger number of SFS loci.

4.2 Filtering out non-recombining loci

Our method requires an important pre-processing step to find loci as contiguous blocks of segregating sites that are free of intra-locus recombination. One could use for example the simple four-gamete test [HK85], or more complex methods for detecting blocks of loci that are free of intra-locus recombination (for eg. [Pos02]). Such filtered loci can then be summarized 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.

4.3 Scalable computing framework

Apache Spark, a unified engine for big data processing [ZXW$^+$16], and the ADAM module [MNH$^+$13] 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 [SV18] 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.

4.4 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 [STH$^+$11], 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 [PVWR18]).

Acknowledgments. The authors would like to thank the two referees for their careful

reading and their numerous comments which helped to improve the presentation of the results.

Data accessibility. The pseudo-code developed for this work is given in the Supplementary Material. It is also publicly shared as sageMath/python code with all examples discussed here [SV18].

Competing interests. We have no competing interests.

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.

References

[Ald01] D.J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16(1):23–34, 2001.

[BEKV13] N.H. Barton, A.M. Etheridge, J. Kelleher, and A. Véber. Inference in two dimensions: allele frequencies versus lengths of shared sequence blocks. *Theor. Pop. Biol.*, 87:105–119, 2013.

[BFL15] L. Bunnefeld, L.A.F. Frantz, and K. Lohse. Inferring bottlenecks from genome-wide samples of short sequence blocks. *Genetics*, 201(3):1157–1169, 2015.

[BRJ+16] S. Boitard, W. Rodriguez, F. Jay, S. Mona, and F. Austerlitz. Inferring population size history from large samples of genome-wide molecular data - An Approximate Bayesian Computation approach. *PLoS Genetics*, 12(3):e1005877, 2016.

[BZB02] M.A. Beaumont, W. Zhang, and D.J. Balding. Approximate Bayesian Computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[DIG04] M. De Iorio and R.C. Griffiths. Importance sampling on coalescent histories. *Adv. Appl. Prob.*, 36:417–433, 2004.

[DM04] Pierre Del Moral. *Feynman-Kac formulae : genealogical and interacting particle systems with applications*. Springer, New York, 2004.

[FD01] P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318, 2001.

[GHWB09] R.N. Gutenkunst, R.D. Hernandez, S.H. Williamson, and C.D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695, 2009.

[GJB13] L.M. Gattepaille, M. Jakobsson, and M.G.B. Blum. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity*, 110(5):409–419, 2013.

[HD08] J. Heled and A.J. Drummond. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, 8:289, 2008.

[HK85] R.R. Hudson and N.L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–164, 1985.

[HN13] K. Harris and R. Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, 9(6):e1003521, 2013.

[HS11] S.Y.W. Ho and B. Shapiro. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Res.*, 11:423–434, 2011.

[Hud02] R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, 2002.

[HUW08] A. Hobolth, M.K. Uyenoyama, and C. Wiuf. Importance sampling for the infinite sites model. *Statistical Applications in Genetics and Molecular Biology*, 7(1):32, 2008.

[KJS15] J. Koskela, P.A. Jenkins, and D. Spano. Computational inference beyond Kingman's coalescent. *J. Appl. Probab.*, 52(2):519–537, 2015.

[KSCS16] J.A. Kamm, J.P. Spence, J. Chan, and Y.S. Song. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. *Genetics*, 203(3):1381–1399, 2016.

[MC05] G. McVean and N. Cardin. Approximating the coalescent with recombination. *Phil. Trans. Royal Soc. B*, 360:1387–1393, 2005.

[MH97] A.O. Mooers and S.B. Heard. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, 72(1):31–54, 1997.

[MNH+13] Matt Massie, Frank Nothaft, Christopher Hartl, Christos Kozanitis, Andr Schumacher, Anthony D. Joseph, and David A. Patterson. ADAM: Genomics formats and processing patterns for cloud scale computing. Technical Report UCB/EECS-2013-207, EECS Department, University of California, Berkeley, Dec 2013.

[MW06] P. Marjoram and J. Wall. Fast "coalescent" simulation. *BMC Genetics*, 7:16, 2006.

[Nie00] R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154:931–942, 2000.

[Nie05] R. Nielsen. Molecular signatures of natural selection. *Annu. Rev. Genet.*, 39:197–218, 2005.

[Pos02] David Posada. Evaluation of methods for detecting recombination from DNA sequences: Empirical data. *Molecular Biology and Evolution*, 19(5):708–717, 2002.

[PRH00] O.G. Pybus, A. Rambaut, and P.H. Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155:1429–11437, 2000.

[PVWR18] J.A. Palacios, A. Véber, J. Wakeley, and S. Ramachandran. BESTT: Bayesian Estimation by Sampling Tajima's Trees. *In preparation*, 2018.

[PWE10] B.M. Peter, D. Wegmann, and L. Excoffier. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol. Ecol.*, 19:4648–4660, 2010.

[PWR15] J.A. Palacios, J. Wakeley, and S. Ramachandran. Bayesian nonparametric inference of population size changes from sequential genealogies. *Genetics*, 201:281–304, 2015.

[RPR+12] C. Roux, M. Pauwels, M.-V. Ruggiero, D. Charlesworth, V. Castric, and X. Vekemans. Recent and ancient signature of balancing selection around the S-locus in *Arabidopsis halleri* and *A. lyrata*. *Mol. Biol. Evol.*, 30(2):435–447, 2012.

[SH92] S.A. Sawyer and D.L. Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132:1161–1176, 1992.

[SKS16] M. Steinrücken, J.A. Kamm, and Y.S. Song. Inference of complex population histories using whole-genome sequences from multiple populations. *BioRxiv preprint*, 2016.

[SP97] Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, Dec 1997.

[SSV15] R. Sainudiin, T. Stadler, and A. Véber. Finding the best resolution for the Kingman-Tajima coalescent: theory and applications,. *J. Math. Biol.*, 70:1207–1247, 2015.

[STH+11] R. Sainudiin, K. Thornton, J. Harlow, J. Booth, M. Stillman, R. Yoshida, R.C. Griffiths, G. McVean, and P. Donnelly. Experiments with the Site Frequency Spectrum. *Bulletin of Mathematical Biology*, 73(4):829–872, 2011.

[SV18] R. Sainudiin and A. Véber. UnfoldingSFS. Technical report, `https://cocalc.com/share/ac7f397f-eab9-45fc-9278-f486af09ca55/FullLikelihoodInferenceSFS.sagews?viewer=share`, 2018.

[Taj83] F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.

[TK10] S.T. Tokdar and R.E. Kass. Importance sampling: a review. *Wiley Interdisc. Rev. Comput. Stat.*, 2:54–60, 2010.

[ZXW+16] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache Spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, October 2016.