The Annals of Applied Statistics
2020, Vol. 14, No. 2, 727–751
https://doi.org/10.1214/19-AOAS1313

SEQUENTIAL IMPORTANCE SAMPLING FOR MULTIRESOLUTION KINGMAN–TAJIMA COALESCENT COUNTING

BY LORENZO CAPPELLO[1,*] AND JULIA A. PALACIOS[1,2,†]

[1]*Department of Statistics, Stanford University*
[2]*Department of Biomedical Data Science, Stanford University,* [*]*cappello@stanford.edu;* [†]*juliapr@stanford.edu*

Statistical inference of evolutionary parameters from molecular sequence data relies on coalescent models to account for the shared genealogical ancestry of the samples. However, inferential algorithms do not scale to available data sets. A strategy to improve computational efficiency is to rely on simpler coalescent and mutation models, resulting in smaller hidden state spaces. An estimate of the cardinality of the state space of genealogical trees at different resolutions is essential to decide the best modeling strategy for a given dataset. To our knowledge, there is neither an exact nor approximate method to determine these cardinalities. We propose a sequential importance sampling algorithm to estimate the cardinality of the sample space of genealogical trees under different coalescent resolutions. Our sampling scheme proceeds sequentially across the set of combinatorial constraints imposed by the data which, in this work, are completely linked sequences of DNA at a nonrecombining segment. We analyze the cardinality of different genealogical tree spaces on simulations to study the settings that favor coarser resolutions. We apply our method to estimate the cardinality of genealogical tree spaces from mtDNA data from the 1000 genomes and a sample from a Melanesian population at the β-globin locus.

1. Introduction. Statistical inference of evolutionary parameters, such as effective population size $N(t)$, from molecular sequence data is an important task in population genetics, conservation biology, anthropology and public health (Liu et al. (2013), Nordborg (1998), Rosenberg and Nordborg (2002)). Inference of such parameters relies on the coalescent process that explicitly models the shared ancestry of a sample (genealogy) of n individuals from a population. More specifically, in the standard neutral coalescent framework observed molecular data $\mathbf{Y}$ at a nonrecombining segment from a sample of n individuals within a population is the result of a point process of mutations with rate μ superimposed on the genealogy $\mathbf{g}$ of the sample. The genealogy itself is not directly observed, but it is assumed to be a realization of a stochastic ancestral process (coalescent process) that depends on $N(t)$. Figure 1 shows a realization of the standard coalescent (genealogy) and mutations.

Both Bayesian and frequentist methods rely on the marginal likelihood calculated by integrating over the latent space of genealogies, that is,

$$P\big(\mathbf{Y}|N(t),\mu\big) = \int_{\mathbf{g}\in\mathcal{G}\times\mathbb{R}^{n-1}} P(\mathbf{Y}\mid \mathbf{g},\mu)P\big(\mathbf{g}\mid N(t)\big)\,d\mathbf{g}. \tag{1.1}$$

Integration in the previous equation involves the sum over all possible tree topologies and $n-1$ integrals over coalescent times $\mathbf{t}\in\mathbb{R}^{n-1}$ (bifurcating times). The integral in (1.1) is usually approximated via Monte Carlo (MC) or Markov chain Monte Carlo (MCMC). However, the cardinality of the hidden state space of tree topologies $|\mathcal{G}|$ grows superexponentially with the number of samples n, making integration over the space of genealogies already challenging for small n.

Received February 2019; revised September 2019.
Key words and phrases. Coalescent, sequential importance sampling, enumeration.

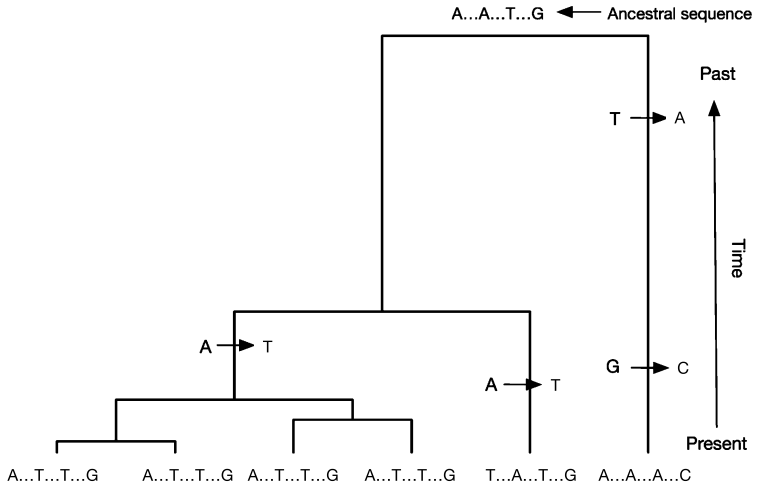

FIG. 1. *Coalescence and mutation. A genealogy of six individuals at a locus of* 100 *base pairs is depicted as a bifurcating tree. Four mutations* (*at different sites*) *are superimposed along the branches of the tree giving rise to the six sequences shown at the tips of the tree. The* 96 *sites* (*base pairs*) *that do not mutate are represented by dots, and only the nucleotides at the polymorphic sites are shown.*

In order to gain computational tractability, several methods have been proposed to infer $N(t)$ from summary statistics, such as the site frequency spectra (Terhorst, Kamm and Song (2017)) from an estimated genealogy (Palacios and Minin (2013), Gattepaille, Günther and Jakobsson (2016)) or from a small number of samples (Drummond et al. (2012)). Gao and Keinan (2016) present an extensive list of implemented methods.

Alternative approaches that rely on lower resolution coalescent models have been recently proposed (Sainudiin, Stadler and Véber (2015), Sainudiin and Véber (2018), Palacios et al. (2019+)). The appealing advantage of these approaches is the a priori drastic reduction in the cardinality of the space of tree topologies for a fixed n. However, conditionally on a given dataset and the popular *infinite sites* mutation model (Watterson (1975)), the true reduction in cardinality, that is, the number of tree topologies for which $P(\mathbf{Y} \mid \mathbf{g}, \mu) > 0$ (compatible) is known neither analytically nor approximately.

In this work we propose a set of algorithms to approximate the cardinality of different tree topology spaces modeled at different coalescent resolutions, the so-called Kingman–Tajima resolutions (Sainudiin, Stadler and Véber (2015)). Reliable estimation of the cardinality of the coalescent hidden state space should provide valuable guidance to statisticians in designing methods employing these different resolutions. State-space count also offers an important auxiliary tool for practitioners: first, it can aid tuning parameters of the MCMC chains, for example, length of the chain; second, and closely related, it is informative of the computational feasibility of coalescent based inference for a given dataset, for example, we will quantify how it is not solely sample size that drives computational feasibility, but for fixed n, the state space size and, consequently, the computational burden, varies largely as a function of the data at hand; lastly, it may offer a convergence diagnostic criteria for sampling methods, for example, what proportion of the state space has been explored in the approximate posterior distribution. In addition, the cardinality of the topological tree space is already an input of some inferential algorithms beyond MCMC, such as the combinatorial sequential Monte Carlo (Wang, Bouchard-Côté and Doucet (2015)).

Counting genealogical trees is a very active area of research in biology and mathematics starting from Cayley (1856); see Steel (2016) for a review. To our knowledge, the large body of work in this area has focused on *exact* combinatorial results or *recursive* algorithms to explore a constrained space. In this work the combinatorial question of counting the number

of compatible tree topologies with the data is treated as a *statistical* problem—estimation of the normalizing constant of a uniform discrete distribution over the space of compatible tree topologies (Jerrum, Valiant and Vazirani (1986)). In this work we estimate the normalizing constant by sampling compatible trees.

Lacking a trivial uniform sampling algorithm in this context, there are two classes of methods to estimate the cardinality of discrete structures subject to constraints, MCMC and sequential importance sampling (SIS). There is a large literature documenting both applications to challenging combinatorial problems and good empirical performances of both MCMC methods (Blanchet and Rudoy (2009), Jerrum and Sinclair (1996), Sinclair (2012)) and SIS methods (Blitzstein and Diaconis (2010), Chen and Chen (2018), Chen et al. (2005), Diaconis (2018), Knuth (1976)). However, there is not a prevailing consensus that one method outperforms the other, even within the same application. Moreover, we are not aware of the use of these methods in the context of coalescent models.

Our estimation method is an instance of SIS. More specifically, our algorithm sequentially samples topologies g compatible with the data with a tractable sampling probability $q(g)$. The SIS estimation of the cardinality is computed by a Monte Carlo approximation of the following expectation:

$$\mathrm{E}_q\left[\frac{1}{q(g)}\right] = \sum_{g \in \mathcal{G}_C} \frac{1}{q(g)} q(g) = |\mathcal{G}_C|, \tag{1.2}$$

where $\mathcal{G}_C$ is the space of compatible tree topologies. The main contribution of this work is a set of algorithms that sample only compatible tree topologies under different coalescent models and, consequently, correspond to different proposals q. Whereas the focus of this work is the estimation of state-space cardinalities, it is easy to see that the same procedure can be applied to enumerate tree topologies with certain features of interest. For example, the number of balanced trees and trees with certain shapes are indicatives of population structure and phylogenetic diversity (Ferretti et al. (2017), Maliet, Gascuel and Lambert (2018)); the number of cherries and pitchforks are indicatives of neutrality (Disanto and Wiehe (2013), Griffiths (1987)). Although we do not explore this research direction in this paper, our algorithms offer a building block to study how a neutral coalescent model fits the data set at hand.

The rest of the paper proceeds as follows. Section 2 reviews the Kingman–Tajima coalescent and the perfect phylogeny representation of molecular sequence data. In Section 3 we present the sampling algorithms. In Section 4 we analyze the cardinality of genealogical spaces under different coalescent resolutions from simulated data, and in Section 5 we present two case studies, one case study from simulated data and one case study of a sample of human mtDNA from the 1000 genomes and other human DNA datasets. Section 6 concludes.

2. Preliminaries.

2.1. *Kingman–Tajima coalescent.* *Kingman's coalescent* is a continuous-time Markov chain with state space the set of partitions of the label set $[n] = \{1, \ldots, n\}$ of the n individuals in a sample (Kingman (1982)). The process starts at $\{\{1\}, \ldots, \{n\}\}$; it then jumps when two of the n individuals coalesce (represented as the merger of two branches in a single internal node in the genealogy). The state of the process after the first transition is the partition of $[n]$ into $n-1$ sets, one set with the labels of the two individuals that coalesce and $n-1$ singleton sets with the labels of the remaining individuals. The process ends when all individuals coalesce, that is, at state $\{1, \ldots, n\}$ when there is a single set (at the root of the genealogy when all individuals have a common ancestor).

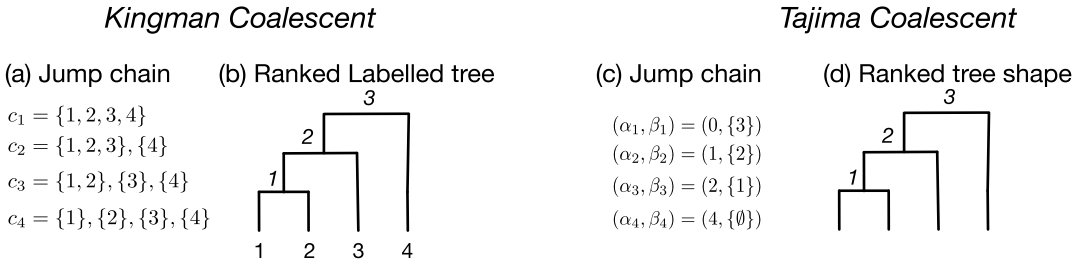

FIG. 2. *Coalescent tree topologies.* (*a*) *A complete realization from Kingman's jump chain and* (*b*) *its corresponding bijection, a ranked labeled tree topology.* (*c*) *A complete realization from Tajima's jump chain and* (*d*) *its corresponding bijection, a ranked tree shape.*

A complete realization of Kingman's coalescent process is commonly represented as a timed bifurcating tree (genealogy) denoted by $\mathbf{g}^K = \{g^K, \mathbf{t}\}$. In this work we concern ourselves with the tree topology only, that is, a complete realization of the embedded jump chain of the process $g^K = \{c_i\}_{i=n:1}$, where c_i is the state of the process when there are i branches. A genealogical representation of g^K is given in Figure 2(b) and the corresponding chain in Figure 2(a). Superindex K in g^K serves to distinguish a Kingman's tree topology to any other type of tree topology. The transition probability of the jump chain is

$$P(c_{i-1} \mid c_i) = \begin{cases} \binom{i}{2}^{-1} & \text{if } c_{i-1} \prec c_i, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where $c_{i-1} \prec c_i$ means that c_{i-1} can be obtained from joining two elements of c_i. It follows from (2.1) that $P(g^K) = 2^{n-1}/[n!(n-1)!]$, that is, the discrete uniform over all possible chain trajectories. We will use $\mathcal{G}_n^K$ to denote the space of such Kingman's topologies.

Tajima's coalescent is a continuous-time Markov chain whose complete realization is also in bijection with a timed bifurcating tree. Its embedded jump chain $\{(\alpha_i, \beta_i)\}_{i=n:1}$ keeps track of the number of singletons α_i and the set of extant vintage labels β_i when there are i branches (Sainudiin, Stadler and Véber (2015), Tajima (1983)). We refer to singleton branch as a branch in the tree that subtends a leaf and a vintage as the internal branch that subtends the subtree labeled by the ranking at which the subtree was created in the jump chain. Since singletons' labels are ignored, there are up to three types transitions: two singletons merge, one singleton and a vintage merge, or two vintages merge. Formally, given a current state (α_j, β_j), when there are $j = \alpha_j + |\beta_j|$ branches in the genealogy, the chain transitions to $\alpha_{j-1} = \alpha_j - 2$ and $\beta_{j-1} = \beta_j \cup \{j\}$ if two singletons create a new vintage node with label $\{j\}$; the chain transitions to $\alpha_{j-1} = \alpha_j - 1$ and $\beta_{j-1} = \beta_j \setminus \{i\} \cup \{j\}$ if one singleton and vintage branch with label $\{i\}$ merge to create a new vintage node with label $\{j\}$, and the chain transitions to $\alpha_{j-1} = \alpha_j$ and $\beta_{j-1} = \beta_j \setminus \{i,k\} \cup \{j\}$ if vintages $\{i\}$ and $\{k\}$ merge to create a new vintage with label $\{j\}$. The process starts at state $\alpha_n = n$ and $\beta_n = \varnothing$ (at the tips of the tree). The chain then jumps to $\alpha_{n-1} = n-2$ and $\beta_{n-1} = \{1\}$ (with probability one since this is the only possible transition at this step), and the vintage $\{1\}$ is created. The process ends at the root when there is a single vintage, that is, $\alpha_1 = 0$, and $\beta_1 = \{n-1\}$. A complete realization of Tajima's coalescent continuous process can be represented as a genealogy $\mathbf{g}^T = \{g^T, \mathbf{t}\}$. A complete realization of the jump chain of the process is denoted by $g^T = \{(\alpha_i, \beta_i)\}_{i=n:1}$ (Figure 2(c)). The jump chain has the following transition probabilities:

$$P\big[(\alpha_{i-1}, \beta_{i-1}) | (\alpha_i, \beta_i)\big] = \begin{cases} \dfrac{\binom{\alpha_i}{\alpha_i - \alpha_{i-1}}}{\binom{\alpha_i + |\beta_i|}{2}} & \text{if } (\alpha_{i-1}, \beta_{i-1}) \prec (\alpha_i, \beta_i), \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

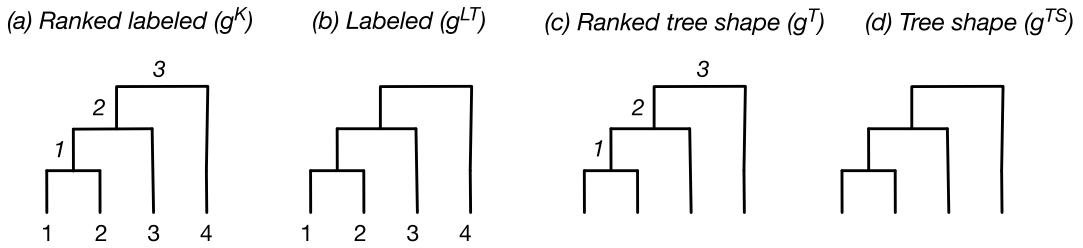

FIG. 3. *Tree topologies*: *The* (*a*) *ranked labeled tree topology* (g^K, *Kingman*), (*b*) *labeled* (*unranked*) *tree topology* (g^{LT}), (*c*) *ranked tree shape* (g^T, *Tajima*) *and* (*d*) *tree shape* (g^{TS}).

Given (2.2), one can compute the probability of a Tajima's tree topology g^T as $P(g^T) = 2^{n-c(g)-1}/(n-1)!$, where $c(g)$ is the number of coalescent events joining two singletons. We will use $\mathcal{G}_n^T$ to denote the space of such Tajima's topologies.

While Kingman's coalescent keeps track of who is related to whom, Tajima's coalescent describes the evolutionary relationships of a sample of n individuals by keeping track of the number of singletons and the vintage labels of extant "families". We note that Tajima's coalescent has the same number of transitions and wait time distribution as Kingman's coalescent. Tajima's coalescent is a lower-resolution coalescent process since it takes values in a smaller state space than Kingman's. Sainudiin, Stadler and Véber (2015) formalize this notion and describe in detail other coalescent resolutions.

The corresponding tree topology under Kingman coalescent g^K is a *ranked labeled tree*, and the corresponding tree topology under Tajima coalescent g^T is a *ranked tree shape* (Figure 2). The formal definitions are as follows:

DEFINITION 1. A *ranked labeled tree* is a rooted binary tree with unique labels at the tips and a total ordering (ranking) for the internal nodes.

DEFINITION 2. A *ranked tree shape* is a rooted binary unlabeled tree with a total ordering (ranking) for the internal nodes.

Although our main objective is to analyze Kingman and Tajima tree topologies, we extend our analysis to the corresponding unranked tree topologies, unranked labeled tree and tree shapes. Figure 3 shows the four tree topologies analyzed in this manuscript.

There are explicit or recursive formulas to compute the number of topologies with n leaves. The number of ranked labeled trees is $|\mathcal{G}_n^K| = n!(n-1)!/2^{n-1}$; the number of unranked labeled trees (binary phylogenetic trees) is $|\mathcal{G}_n^{\mathrm{LT}}| = (2n-3)!!$ (Steel (2016)); the number of ranked tree shapes $|\mathcal{G}_n^T|$ is the nth term of the Euler zig-zag sequence (alternating permutations, OEIS: A000111) (Disanto and Wiehe (2013)), and the number of tree shapes is the nth Wedderburn–Etherington number (OEIS: 01190) (Steel (2016)).

For $n > 3$, it holds that $|\mathcal{G}_n^{\mathrm{TS}}| < |\mathcal{G}_n^{\mathrm{LT}}|$ and $|\mathcal{G}_n^T| < |\mathcal{G}_n^K|$, that is, the unlabeled tree topologies have smaller cardinalities that the labeled counterparts. For example, for $n = 5$, there are 180 ranked labeled trees and *five* unlabeled ranked trees. Similarly, 105 labeled trees and *three* tree shapes. This cardinality difference has motivated the study of lower resolution coalescent processes (Sainudiin, Stadler and Véber (2015)). However, it is not clear how big this difference is when the observed data restricts the space of topologies. In the next section we describe how observed data imposes combinatorial constraints on the topological space.

2.2. *Perfect phylogeny and infinite sites model*. As mentioned in the Introduction, we assume that molecular variation at a nonrecombining contiguous segment of DNA (or locus) is the result of a mutation process superimposed on the timed genealogy $\mathbf{g}$ (Figure 1).

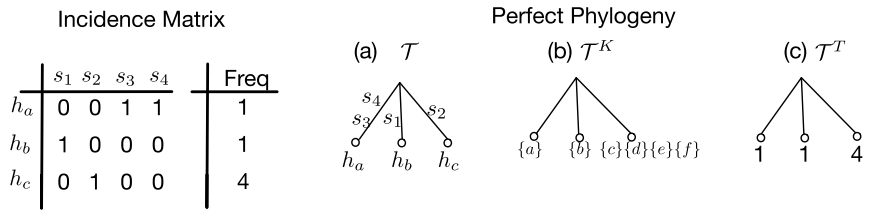

FIG. 4. *Incidence matrix and perfect phylogeny representation. Data is summarized as an incidence matrix* (h *denotes the haplotypes*, s *the segregating sites*) *and a vector of frequencies.* (*a*) *Original perfect phylogeny* $\mathcal{T}$ *in bijection with the incidence matrix*; *each of the four polymorphic sites labels exactly one edge. When an edge has multiple labels, the order of the labels is irrelevant. Each of the three haplotypes labels one leaf if* $\mathcal{T}$. (*b*) *Kingman's perfect phylogeny* $\mathcal{T}^K$: *it is a perfect phylogeny with edge labels removed and leaf labels the set of individual labels for each haplotype.* (*c*) *Tajima's perfect phylogeny* $\mathcal{T}^T$: *it is a perfect phylogeny with edge labels removed and leaf labels the corresponding haplotype frequency.*

Here, we assume that mutations (or substitutions) occur at sites that have not mutated previously. This mutation model is called the *infinite-sites model* (ISM) (Kimura (1969), Watterson (1975)). Further, we assume that our data consists of a single nonrecombining segment of DNA. Although we will not model the mutation process explicitly, it is commonly assumed that mutation happens as Poisson process on the timed genealogy $\mathbf{g}$. However, an important consequence is that the ISM imposes a restriction on the space of tree topologies: given that at most one mutation occurs at a site, this mutation must occur on a branch subtending individuals with the observed mutation (Figure 1). Therefore, mutations partition the observed sequences into two sets: the sequences that carry the mutations and the sequences that do not. In addition, if the ancestral type at each polymorphic site is known, molecular data from n individuals at m polymorphic sites can be represented as an incidence matrix $\mathbf{Y}$ and a vector of the row frequencies of the matrix $\mathbf{Y}$. The incidence matrix $\mathbf{Y}$ is a $k \times m$ matrix with 0–1 entries, where 0 indicates the ancestral type and 1 indicates the mutant type; k is the number of unique sequences (or haplotypes) observed in the sample, and the vector of frequencies indicates the number of times each haplotype is observed in the sample. For example, the $n = 6$ sequences displayed at the leaves of the genealogy in Figure 1 can be summarized as the incidence matrix and corresponding frequency vector in Figure 4. The three haplotypes in this example are A...A...A...C, T...A...T...G and A...T...T...G with labels h_a, h_b and h_c, respectively. In this example, the ancestral sequence is displayed at the root of the tree in Figure 1. In what follows, we will assume that our data are an incidence matrix and corresponding frequencies as in Figure 4.

Gusfield (1991) proposed an algorithm to represent the incidence matrix as a multifurcating tree called *perfect phylogeny*. A perfect phylogeny is in bijection with an incidence matrix, and it exists if and only if the infinite sites and no recombination assumptions hold. In our example, the multifurcating tree displayed in Figure 4(a) is the corresponding perfect phylogeny representation of the incidence matrix. The key in the perfect phylogeny representation is that mutations (labeled as $s_1, \ldots, s_4$ in Figure 4) partition the haplotypes into different groups (three groups represented as leaf nodes in Figure 4(a)) and thus enforce a combinatorial constraint.

More formally, given an incidence matrix $\mathbf{Y}$, a *perfect phylogeny* $\mathcal{T}$ is a rooted tree (possibly multifurcating) with k leaves and satisfying the following properties:

1. Each of the k haplotypes labels one leaf in $\mathcal{T}$.
2. Each of the m polymorphic sites labels exactly one edge. When multiple sites label the same edge, the order of the labels along the edge is arbitrary. Some external edges (edges subtending leaves) may not be labeled, indicating that they do not carry additional mutations to their parent node.

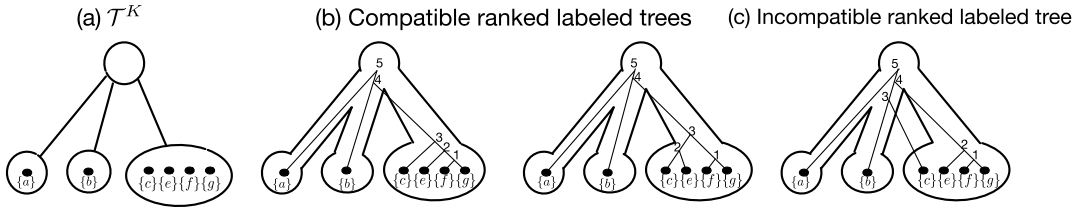

FIG. 5. *Compatibility of ranked labeled trees with the perfect phylogeny.* (*a*) *Kingman perfect phylogeny*, (*b*) *two examples of ranked labeled trees compatible with the perfect phylogeny*, (*c*) *incompatible ranked labeled tree* ({*c*} *should coalesce first with individuals in its node*).

3. For any haplotype h_k, the labels of the edges along the unique path from the root to the leaf h_k specify all the sites where h_k has the mutant type.

Allow a few remarks. The tree $\mathcal{T}$ is usually not the tree topology of a coalescent genealogy. First, each leaf node labels a unique haplotype that could have been sampled with frequency higher than one. Second, we have restricted our attention to binary trees, those sampled from one of the coalescent processes and $\mathcal{T}$ is not necessarily binary (in most cases it is not).

To simplify our exposition in the following sections, we summarize the perfect phylogeny somewhat different than the original Gusfield's algorithm depending on whether we wish to count Kingman's or Tajima's topologies compatible with the observed data. Our perfect phylogeny representation for counting Kingman's tree topologies is denoted by $\mathcal{T}^K$. In $\mathcal{T}^K$, we remove the edge labels and label the leaf nodes by the set of individuals labels that share the same haplotype. For example, in Figure 4(b) individuals $\{c\}$, $\{d\}$, $\{e\}$ and $\{f\}$ share the same haplotype h_c. In the case when a haplotype leaf descends from an edge with no mutations, we attach the set of individuals labels to its parent node and remove the leaf. Similarly, our perfect phylogeny representation for counting Tajima's tree topologies is denoted by $\mathcal{T}^T$. In $\mathcal{T}^T$, we again remove edge labels, but now we label leaf nodes by the frequency of their corresponding haplotypes (Figure 4(c)). Note that such a representation reflects the fact that two individuals sharing the same mutations are indistinguishable. In the case when a haplotype leaf descends from an edge with no mutations, we attach the frequency of the haplotype to its parent node and remove the leaf.

A tree topology g is *compatible* with the perfect phylogeny $\mathcal{T}$ if $P(\mathcal{T}|g, \mathbf{t}) > 0$. That is, if all sequences descending from a node V in $\mathcal{T}$ coalesce in g before coalescing with any other sequence descending from a different node U in $\mathcal{T}$. Figure 5(b) shows examples of two compatible ranked labeled trees with the perfect phylogeny in Figure 4(b) and 5(a), while Figure 5(c) shows an incompatible ranked labeled tree topology. The topology in Figure 5(c) is not compatible since there is no node in g^K that groups together $\{c\}$, $\{e\}$, $\{f\}$, $\{g\}$ without $\{a\}$ or $\{b\}$. In the following sections we describe our algorithms for approximating the number of tree topologies compatible with a given perfect phylogeny. In the following we denote the set of compatible tree topologies by $\mathcal{G}_{n,C} \subseteq \mathcal{G}_n$.

3. Sequential importance sampling. Let p denote the uniform discrete distribution on $\mathcal{G}_{n,C}$. Suppose we can sample from a distribution q with support $\mathcal{G}_{n,C}$, then the normalizing constant of p, that is, $|\mathcal{G}_{n,C}|$ is given by

$$\mathrm{E}_q\left[\frac{1}{q(g)}\right] = \sum_{g \in \mathcal{G}_{n,C}} \frac{1}{q(g)} q(g) = |\mathcal{G}_{n,C}|, \tag{3.1}$$

which, given an $i.i.d.$ sample from q of size N, can be approximated via Monte Carlo by

$$\widehat{|\mathcal{G}_{n,C}|} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{q(g_i)} \tag{3.2}$$

with standard error: $\mathrm{se}(\widehat{|\mathcal{G}_{n,C}|}) = \sqrt{\mathrm{Var}_q(1/q(g))}/\sqrt{N}$, and the variance can be approximated with its empirical counterpart.

Average (3.2) is an instance of importance sampling (IS) (Hammersley and Handscomb (1965), Owen (2013)). As described in previous sections, observed data impose combinatorial constraints to the space of compatible tree topologies. The idea is to construct a compatible tree topology $g \in \mathcal{G}_{n,C}$ sequentially with choices $c_n, \ldots, c_1$ (one coalescence at a time) from the tips to the root, ensuring that each choice is compatible with the observed data (or perfect phylogeny) and with known probability

$$q(g) = q(c_n)q(c_{n-1} \mid c_n) \cdots q(c_1 \mid c_2). \tag{3.3}$$

Approaches with a similar stochastic sequential nature construction have been used for enumeration in other contexts, such as random graphs, networks and contingency tables (Blitzstein and Diaconis (2010), Chen and Chen (2018), Chen et al. (2005), Knuth (1976), Diaconis (2018)).

It is clear from this literature that the algorithm should satisfy two desiderata: it should not "get stuck," that is, it should not sample g outside $|\mathcal{G}_{n,C}|$; in addition, $q(g)$ should be easily computed.

How large N should be largely depends on how close the proposal distribution q is to the target distribution p. In our problem, p is uniform discrete on the set of compatible trees.

Chatterjee and Diaconis (2018) show that $N \approx \exp(\mathrm{KL}(q, p))$ is necessary and sufficient for accurate estimation by IS, where KL denotes the Kullback–Leibler divergence. In addition, Chatterjee and Diaconis (2018) warn against the use of sample variance as a criterion for IS convergence; they prove that it can be arbitrary small for large N independently from p and q.

A common metric to assess convergence is the importance sampling effective sample size ESS, where $\mathrm{ESS} = N/(1 + \mathrm{cv}^2)$ and cv^2 is the coefficient of variation given by

$$\mathrm{cv}^2 = \frac{\mathrm{Var}_q[p(g)/q(g)]}{\mathrm{E}_q^2[p(g)/q(g)]}$$

and estimated empirically. cv^2 is the χ^2-distance between p and q. A low cv^2 (ESS close to N) is a good indicator of the quality of the proposal q.

In lieu of sample variance as a metric for convergence, Chatterjee and Diaconis (2018) define $q_N = \mathrm{E}[Q_N]$ where

$$Q_N = \frac{\max_{1 \le i \le N} p(g_i)/q(g_i)}{\sum_{i=1}^N p(g_i)/q(g_i)}$$

and propose to use a Monte Carlo estimate of q_N below a certain threshold as a criterion for convergence. A low value of q_n can be interpreted as a situation in which a sufficiently large number of samples have been collected (large denominator) to counterbalance the effect of possible "outliers" that are sampled (large numerator). Computing a Monte Carlo estimate is computationally expensive and, hence, in this work we simply compute a single running Q_N and combine it with the other metrics described. Note that, since we restrict our attention to p uniform discrete, the normalizing constant cancels out in both Q_N and cv^2; so it is possible to compute these two diagnostics.

3.1. *Sampling trees compatible with a perfect phylogeny.* To generate a tree topology $g \in \mathcal{G}_{n,C}$ compatible with the observed data $\mathcal{T}$, we proceed sequentially from tips to the root in both $\mathcal{T}$ and g: one coalescence in g and one node in $\mathcal{T}$ at a time. In every step we keep an active set of nodes of $\mathcal{T}$ in which we can sample particles to coalesce. Initially, this set

includes all nodes with at least two particles. We then randomly select an active node in $\mathcal{T}$ and randomly select two particles from the selected node to coalesce in g. At this time, the two selected particles are replaced by a new particle in the selected node in $\mathcal{T}$. If a node in $\mathcal{T}$ has a single particle, the node is removed and its particle is transferred to its parent node. The algorithm ends when $\mathcal{T}$ has a single node with a single particle and when a complete genealogy is generated. All particles in one node must coalesce with each other before they can coalesce with any other particle.

We propose two new algorithms that share the steps just described: one for sampling ranked labeled trees (Kingman trees) and one for sampling ranked tree shapes (Tajima trees). A simple combinatorial argument allows extending the outputs of these two algorithms to their respective unranked counterparts. This extension should be considered by all means a byproduct of the Kingman and Tajima algorithms. It is of interest because we can obtain estimates for two other resolutions at almost no additional computational cost.

We start with some notations. We use V to denote the set of nodes of the perfect phylogeny $\mathcal{T}$ and $L \subset V$ to denote the set of active nodes, that is, nodes with at least two particles; v is an element of V, and $\mathrm{pa}(v)$ denotes the parent node of v (if v is not the root). We use the word particle to refer to individual singletons, elements of a partition of $[n]$ or vintages. Each node in $\mathcal{T}$ has either no particles or a given number of particles assigned (labeled or not). Given n individuals, the $n-1$ iterations required to sample a tree topology are indexed in reverse order, that is, from $n-1$ to 1, to be consistent with the notations used in the jump chains of the n-coalescent. This notation allows us to keep track of how many individuals have yet to coalesce.

We saw that Kingman n-coalescent jump chain induces a uniform distribution with support $\mathcal{G}_n^K$. Given that our target distribution for the Kingman topology is uniform on $\mathcal{G}_{n,C}^K$, we mimic within each node the transition probability of the underlying coalescent jump chain. The active node is sampled with probability proportional to the number of assigned particles. Although Tajima's jump chain does not induce a uniform distribution on $\mathcal{G}_n^T$, it is quite close to being uniform: it is uniform across ranked tree shapes with the same number of cherries.

3.1.1. *Data constrained Kingman coalescent.* To sample a ranked labeled tree $g^K = \{c_i\}_{i=n:1}$ of n individuals compatible with the observed perfect phylogeny $\mathcal{T}^K$, we start at $c_n = \{\{1\}, \ldots, \{n\}\}$. Each leaf node of $\mathcal{T}^K$ defines a partition of c_n, and we use c_n^v to denote the set of particles in node v. An exception occurs when an edge with no mutations subtends a leaf; in this case the parent node gets assigned the particles of the leaf and the leaf is removed.

The first step is to define the set L as the set of nodes with at least two particles. If a node has a single particle ($|c_n^v| = 1$), we transfer the particle to its parent node. Then for each iteration $i = n-1, \ldots, 1$, we sample a node in L with probability proportional to the number of particles in that node: at iteration i, the probability of choosing node $v_i \in L$ is $q(v_i) = |c_{i+1}^{v_i}| / \sum_{j \in L} |c_{i+1}^j|$. The transition from $c_{i+1}^{v_i}$ to $c_i^{v_i}$ consists in joining two particles of $c_i^{v_i}$ uniformly at random. If a node is not sampled, we assume $c_i^v = c_{i+1}^v$. This choice mimics the jump chain of a Kingman n-coalescent; the difference is that the Markov chain moves one step on a constrained state space: $c_i^{v_i}$ in lieu of c_i; that is, the coalescent event in node v_i has probability

$$q\big(c_i^{v_i} \mid c_{i+1}^{v_i}\big) = \begin{cases} \binom{|c_{i+1}^{v_i}|}{2}^{-1} & \text{if } c_i^{v_i} \prec c_{i+1}^{v_i}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

Note that at every iteration $c_i = \bigcup_v c_i^v$. The two probabilities $q(v_i)$ and $q(c_i^{v_i} \mid c_{i+1}^{v_i})$ are all we need to compute the transition probability

$$q(c_i \mid c_{i+1}) = q(v_i) q\big(c_i^{v_i} \mid c_{i+1}^{v_i}\big),$$

Algorithm 1 Sequential sampling on a constrained Kingman tree topology

Inputs: $\mathcal{T}^K$ with c_n^v subsets of singletons at all nodes with particles and $c_n^v = \varnothing$ at all remaining nodes.
Outputs: $g^K, q(g^K)$

1. If a node v is such that $|c_n^v| = 1$, then we let $c_n^{\mathrm{pa}(v)} = c_n^{\mathrm{pa}(v)} \cup c_n^v$ and $c_n^v = \varnothing$.
2. Define L as the list of nodes such that $|c_n^v| > 1$
3. Initialize $q = 1$
4. **for** $i = n - 1$ to 1 **do**
   (a) Sample node v_i in L with probability $q(v_i)$.
   (b) Choose particles in v_i to coalesce with probability $q(c_i \mid c_{i+1})$.
   (c) Update $c_i^{v_i}$ and define $c_i^v = c_{i+1}^v$ for all the other nodes.
   (d) If $|c_i^{v_i}| = 1$, we let $c_i^{\mathrm{pa}(v_i)} = c_i^{\mathrm{pa}(v_i)} \cup c_i^{v_i}$ and $c_{i-1}^{v_i} = \varnothing$.
   (e) Update $q = q \times q(v_i) \times q(c_i \mid c_{i+1})$
   (f) Update L as the list of nodes such that $|c_i^v| > 1$
5. **end for**

where $c_i = c_{i+1} \setminus c_{i+1}^v \cup c_i^v$ can be constructed recursively. The last iteration happens at the root node of $\mathcal{T}^K$ and $q(g^T)$ is computed as the product of the transition probabilities as in (3.3). We outline our sampling algorithm with the following example and provide the pseudocode in Algorithm 1.

EXAMPLE 1. Consider the perfect phylogeny $\mathcal{T}^K$ in Figure 6(a). To avoid confusion between the nodes' sampling order $(v_{n-1}, \ldots, v_1)$ and node labels, we label the root node j_0 and the leaf nodes j_1, j_2 and j_3. Figure 6 gives a graphical representation of a single run of the algorithm, where one particle is assigned to j_1, one to j_2 and four to j_3. We start with $c_6^{j_1} = \{a\}$, $c_6^{j_2} = \{b\}$, $c_6^{j_4} = \{\{c\}, \{d\}, \{e\}, \{f\}\}$ and $c_6^{j_0} = \varnothing$. Now, both j_1 and j_2 have a single particle: we transfer their particles to the root node and update $c_n^{j_0} = \{\{a\}, \{b\}\}$ (Figure 6(a–b)). The set of active nodes is $L = \{j_0, j_3\}$. At iteration $i = 5$ (first iteration) suppose we sample node $v_5 = j_3$, this happens with probability $4/6$; then d and f coalesce with probability $1/6$ (Figure 6(b)). We update $c_5^{j_3} = \{\{c\}, \{e\}, \{d, f\}\}$. The set of active sample nodes remains $L = \{j_0, j_3\}$. Figure 6(c–f) shows the remaining iterations. The sequence of sampled nodes is $\{v_5 = j_3, v_4 = j_3, v_3 = j_0, v_2 = j_3, v_1 = j_0\}$ with sampling probabil-

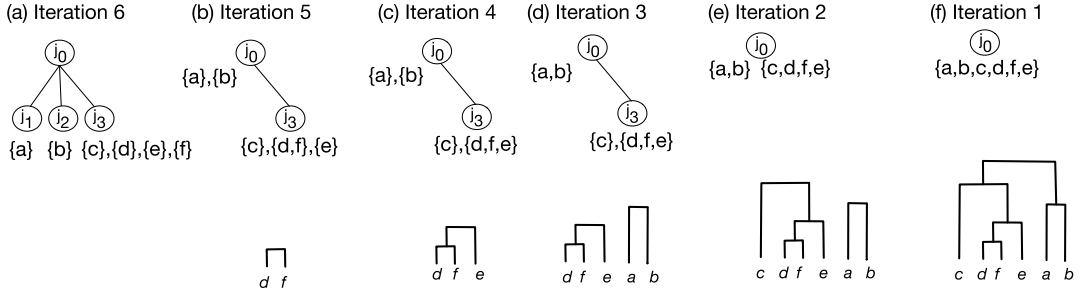

FIG. 6. *Example* 1: *Sequential sampling of a Kingman tree topology with constraints. First row describes the steps in the perfect phylogeny*; *second row describes how* g^K *is sequentially sampled. We start with a perfect phylogeny* (*a*); *in* (*b*) *we assign the singletons to their parent node. In this case* $\{a\}$ *and* $\{b\}$ *are assigned to* j_0. *At each iteration* (*b*)–(*f*) *we select a node and coalesce a pair from the selected node. The algorithm terminates when a single tree topology of size n is generated.*

ities $(4/6, 3/5, 1/2, 1, 1)$. The coalescent events probabilities are $(1/6, 1/3, 1, 1, 1)$. Thus, $q(g^K) = 1/90$.

3.1.2. *Data constrained Tajima coalescent.* To sample a ranked tree shape $g^T = \{(\alpha_i, \beta_i)\}_{i=n:1}$ of n individuals compatible with the observed perfect phylogeny $\mathcal{T}^T$ (Figure 4(c)), we start at $(n, \varnothing)$, and each leaf node in the perfect phylogeny $\mathcal{T}^T$ is assigned a vector (α_n^v, β_n^v). Again, an exception occurs when an edge with no mutations subtends a leaf. In this case we assign the particles to their parent node and remove the leaf node. Recall that α_n^v denotes the number of singletons and β_n^v denotes the set of vintages associated to node v. Initially, each leaf node (possibly also some internal nodes when the no-mutations case occurs) in the perfect phylogeny contains the number of singleton particles $\sum_{v\in V} \alpha_n^v = n$ and no vintages, that is, $\beta_n^v = \varnothing$ for all $v \in V$. At any given iteration i, the number of particles associated to a node v is $\alpha_i^v + |\beta_i^v|$.

Tajima's sampler follows the rationale used to build the Kingman sampler. We define the set L as in the Kingman's sampler (nodes with at least two particles). Then for $n-1$ iterations, we first sample a node $v \in L$ with probability $q(v_i) = (\alpha_i^v + |\beta_i^v|)/\sum_{j\in L}(\alpha_i^j + |\beta_i^j|)$; then, we sample a pair of particles in the selected node to coalesce. Our proposal probability is

$$
\begin{aligned}
&q\big[(\alpha_i^{v_i}, \beta_i^{v_i})|(\alpha_{i+1}^{v_i}, \beta_{i+1}^{v_i})\big] \\
&\quad= \begin{cases} \dbinom{\alpha_{i+1}^{v_i}}{\alpha_{i+1}^{v_i} - \alpha_i^{v_i}} \dbinom{\alpha_{i+1}^{v_i} + |\beta_{i+1}^{v_i}|}{2}^{-1} & \text{if } (\alpha_i^{v_i}, \beta_i^{v_i}) \prec (\alpha_{i+1}^{v_i}, \beta_{i+1}^{v_i}), \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.5}
$$

Analogously to the Kingman sampler, each iteration ends by updating $(\alpha_i^{v_i}, \beta_i^{v_i})$ and L. The pseudocode is presented in Algorithm 2. Note that, as opposed to the Kingman sampler, $q(v_i)$ and $q[(\alpha_i^{v_i}, \beta_i^{v_i})|(\alpha_{i+1}^{v_i}, \beta_{i+1}^{v_i})]$ in Tajima sampling do not fully determine $q[(\alpha_i, \beta_i)|(\alpha_{i+1}, \beta_{i+1})]$, where (α_i, β_i) is the ith state independent of which node is selected; it can be computed as $(\alpha_i, \beta_i) = (\sum_{v\in V} \alpha_i^v, \bigcup_{v\in V} \beta_i^v)$. A transition from $(\alpha_{i+1}, \beta_{i+1})$ to (α_i, β_i) can be obtained by sampling different nodes in the active set, possibly with different sampling probabilities. For example, suppose we are joining two singletons: any $v \in L$ with at least two singletons allows this type of transition. This issue was not relevant in the Kingman sampler because individuals were labeled. Therefore, the output of the sampling algorithm after $n-1$ iterations is $\{(\alpha_i, \beta_i)\}_{i=n:1} = g^T$ along with the sequence of sampling nodes $\boldsymbol{v} = (v_{n-1}, \ldots, v_1)$. It is possible to sample the same g^T with different $\boldsymbol{v}$ and $\boldsymbol{v'}$. These two outputs of the algorithm, which we denote by $(g^T, \boldsymbol{v})$ and $(g^T, \boldsymbol{v'})$, may also have different sampling probabilities $q(g^T, \boldsymbol{v})$ and $q(g^T, \boldsymbol{v'})$. We illustrate this situation with the following example.

EXAMPLE 2. Consider the perfect phylogeny in Figure 7(a). Figure 7(b)–(c) show two ranked tree shapes, g^T and g^{*T}, that can be sampled with our algorithm. Let us

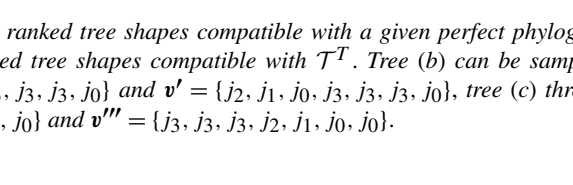

FIG. 7. *Example* 2: *two ranked tree shapes compatible with a given perfect phylogeny.* (*a*) *perfect phylogeny* (*b*)–(*c*) *two possible ranked tree shapes compatible with* $\mathcal{T}^T$. *Tree* (*b*) *can be sampled through two node orderings* $\boldsymbol{v} = \{j_1, j_2, j_0, j_3, j_3, j_3, j_0\}$ *and* $\boldsymbol{v'} = \{j_2, j_1, j_0, j_3, j_3, j_3, j_0\}$, *tree* (*c*) *through four orderings* $\boldsymbol{v}$, $\boldsymbol{v'}$, $\boldsymbol{v''} = \{j_3, j_3, j_3, j_1, j_2, j_0, j_0\}$ *and* $\boldsymbol{v'''} = \{j_3, j_3, j_3, j_2, j_1, j_0, j_0\}$.

first consider g^T in Figure 7(b). A possible sequence of sampling nodes in $\mathcal{T}^T$ is $\boldsymbol{v} = \{j_1, j_2, j_0, j_3, j_3, j_3, j_0\}$. In this case the output of Algorithm 2 would be $(g^T, \boldsymbol{v})$. Although, the sequence $\boldsymbol{v}' = \{j_2, j_1, j_0, j_3, j_3, j_3, j_0\}$ leads also to the same g^T. The two node orderings $\boldsymbol{v}$ and $\boldsymbol{v}'$ can be easily identified in $\mathcal{T}^T$ since nodes j_1 and j_2 are indistinguishable by being siblings of the same size. Let us now turn to g^{*T} in Figure 7(c). In this case there are four possible sampling nodes orderings: $\boldsymbol{v}$, $\boldsymbol{v}'$, $\boldsymbol{v}'' = \{j_3, j_3, j_3, j_1, j_2, j_0, j_0\}$ and $\boldsymbol{v}''' = \{j_3, j_3, j_3, j_2, j_1, j_0, j_0\}$.

We now introduce some notation to distinguish between the output of our sampling algorithm and the elements needed in the sequential importance sampling estimation of $|\mathcal{G}_{n,c}^T|$.

DEFINITION 3. Let $\mathcal{Y}_{n,C}^T$ be the set of all possible outcomes $(g^T, \boldsymbol{v})$ of the Tajima algorithm (Algorithm 2) conditionally on a given perfect phylogeny $\mathcal{T}^T$. We call two outputs of the algorithm: $(g^T, \boldsymbol{v})$ and $(g^T, \boldsymbol{v}')$ equivalent if they have the same ranked tree shape g^T. Define $c^T(g^T)$ as the size of the equivalence class, that is, the number of possible pairs $(g^T, \boldsymbol{v}') \in \mathcal{Y}_{n,C}^T$ equivalent to $(g^T, \boldsymbol{v})$.

It is still possible to use sequential importance sampling despite the fact that our proposal q has support $\mathcal{Y}_{n,C}^T$ instead of $\mathcal{G}_{n,C}^T$. We discuss two alternative ways. The first one is to generate a sample $(g^T, \boldsymbol{v}) \in \mathcal{Y}_{n,C}^T$ with sampling probability $q(g^T, \boldsymbol{v})$ computed as the product of all transition probabilities (Algorithm 2). We then call a backtracking algorithm that lists all possible sequence of nodes $\boldsymbol{v}'$ that would give rise to the same g^T and compute

$$q(g^T) = \sum_{\boldsymbol{v}':(g^T,\boldsymbol{v}')\in\mathcal{Y}_{n,C}^T} q(g^T, \boldsymbol{v}'). \tag{3.6}$$

Finally, we estimate the cardinality of our constrained space by the Monte Carlo approximation to the following:

$$\begin{aligned} E_{\mathcal{Y}_{n,C}^T}\left[\frac{1}{q(g^T)}\right] &= \sum_{(g^T,\boldsymbol{v})\in\mathcal{Y}_{n,C}^T} \frac{q(g^T,\boldsymbol{v})}{q(g^T)} = \sum_{g^T\in\mathcal{G}_{n,C}^T} \frac{1}{q(g^T)} \sum_{\boldsymbol{v}:(g^T,\boldsymbol{v})\in\mathcal{Y}_{n,C}^T} q(g^T,\boldsymbol{v}) \\ &= \sum_{g^T\in\mathcal{G}_{n,C}^T} \frac{q(g^T)}{q(g^T)} = |\mathcal{G}_{n,C}^T|. \end{aligned} \tag{3.7}$$

Note that the backtracking algorithm adds a computational burden to the procedure. The complexity cannot be uniquely determined and, as it is known in the backtracking literature, may vary largely from problem to problem (Knuth (2018)). Since the complexity depends both on the data $\mathcal{T}^T$ and g^T, an analytical expression is not available. We will study this computational burden through simulations in Section 4.

An alternative to the backtracking step is desirable but currently still an open problem. A potential alternative is inspired by a similar situation discussed in Blitzstein and Diaconis (2010) in the context of sampling graphs with a given degree sequence. The cardinality is estimated by the Monte Carlo approximation to the following:

$$\begin{aligned} E_{\mathcal{Y}_{n,C}^T}\left[\frac{1}{c^T(g^T)q(g^T,\boldsymbol{v})}\right] &= \sum_{(g^T,\boldsymbol{v})\in\mathcal{Y}_{n,C}^T} \frac{q(g^T,\boldsymbol{v})}{c^T(g^T)q(g^T,\boldsymbol{v})} \\ &= \sum_{g^T\in\mathcal{G}_{n,C}^T} \frac{1}{c^T(g^T)} \sum_{\boldsymbol{v}:(g^T,\boldsymbol{v})\in\mathcal{Y}_{n,C}^T} 1 = |\mathcal{G}_{n,C}^T|, \end{aligned} \tag{3.8}$$

Algorithm 2 Sampling on the constrained Tajima Space

Inputs: $\mathcal{T}^T$, with α_n^v number of singletons at all leaf nodes (plus the parent node if the leaf subtends from a branch with no mutations), and $\beta_n^v = \varnothing$ for all $v \in V$.
Outputs: g^T, $q(g^T)$

1. If a node v is such that $\alpha_n^v = 1$, then let $\alpha_n^{\mathrm{pa}(v)} = \alpha_n^{\mathrm{pa}(v)} + 1$, $\alpha_n^v = 0$.
2. Define L as the list of nodes with $\alpha_n^v > 1$.
3. **for** $i = n-1$ to 1 **do**:
   (a) Sample node v_i with probability $q(v_i)$.
   (b) Choose particles to coalesce with probability $q[(\alpha_i^{v_i}, \beta_i^{v_i})|(\alpha_{i+1}^{v_i}, \beta_{i+1}^{v_i})]$
   (c) Update $(\alpha_i^{v_i}, \beta_i^{v_i})$ and define $(\alpha_i^{v_i}, \beta_i^{v_i}) = (\alpha_{i+1}^{v_i}, \beta_{i+1}^{v_i})$ for all other nodes
   (d) If $\alpha_i^{v_i} + |\beta_i^{v_i}| = 1$, then let $\alpha_i^{\mathrm{pa}(v_i)} = \alpha_i^{\mathrm{pa}(v_i)} + \alpha_i^{v_i}$, $\alpha_i^{v_i} = 0$, and $\beta_i^{\mathrm{pa}(v_i)} = \beta_i^{\mathrm{pa}(v_i)} \cup \beta_i^{v_i}$, $\beta_i^{v_i} = \varnothing$
   (e) Update $q = q \times q(v_i) \times q[(\alpha_i^{v_i}, \beta_i^{v_i})|(\alpha_{i+1}^{v_i}, \beta_{i+1}^{v_i})]$
   (f) Update L as the list of nodes such that $\alpha_i^v + |\beta_i^v| > 1$.
4. **end for**
5. Compute all possible node paths $\mathbf{v}$ that lead to g^T (backtracking algorithm).
6. Compute $q(g^T)$ as in (3.6).

where $c^T(g^T)$ is the size of the equivalence class $c^T(g^T) = \#\{\boldsymbol{v}' : (g^T, \boldsymbol{v}') \in \mathcal{Y}_{n,C}^T\}$ as in Definition 3. Given a pair $(g^T, \boldsymbol{v})$, we can calculate $c^T(g^T)$ by finding equivalence classes of certain subtrees in g^T relative to $\mathcal{T}^T$. Although a practical implementation is computationally prohibitive, we introduce this idea because in the next section we use it to obtain unranked tree topologies (labeled trees and tree shapes) algorithms as a byproduct of the Kingman and Tajima algorithms.

3.1.3. *Labeled trees and tree shapes.* We now turn to the unranked versions, labeled trees and tree shapes. As before, we define equivalence relations that partitions the spaces $\mathcal{G}_{n,C}^K$ and $\mathcal{G}_{n,C}^T$ into equivalence classes that ignore rankings. We show two simple formulas to compute the size of these classes. As opposed to ranked tree shapes, these formulas are easy to implement and allow to build a SIS procedure to estimate $|\mathcal{G}_{n,C}^{\mathrm{LT}}|$ and $|\mathcal{G}_{n,C}^{\mathrm{TS}}|$ using outputs from the Kingman and Tajima algorithms (Algorithm 1 and 2). First, we define the following two equivalence relations and their cardinalities:

DEFINITION 4. For any element $g^K \in \mathcal{G}_{n,C}^K$, let $\mathrm{LT}(g^K)$ denote the corresponding unranked labeled tree $g^{\mathrm{LT}} \in \mathcal{G}_{n,C}^{\mathrm{LT}}$, obtained by removing the rankings from internal nodes of g^K. We call g^K and g'^K equivalent if $\mathrm{LT}(g^K) = \mathrm{LT}(g'^K)$, and we denote the size of the equivalence class by $c^{\mathrm{LT}}(g^K)$.

PROPOSITION 1. *Let $g^K \in \mathcal{G}_{n,C}^K$, and let $g_{i,1}^K$ and $g_{i,2}^K$ be the two subtrees (or clades) that merge at the ith coalescent event for $i = 1, \ldots, n-1$. Then,*

$$c^{\mathrm{LT}}(g^K) = \prod_{i=1}^{n-1} \frac{(|g_{i,1}^K| + |g_{i,2}^K| - 2)!}{(|g_{i,1}^K| - 1)!(|g_{i,2}^K| - 1)!},$$

where $|g_{i,j}^K|$ denotes the number of leaf nodes of $g_{i,j}^K$.

PROOF. Note that $|g_{i,j}^K| - 1$ is the number of coalescent events in subtree $g_{i,j}^K$. For each fixed i, we are computing the number of possible permutations of $(|g_{i,1}^K| + |g_{i,2}^K| - 2)$ coa-

lescent events of elements of two groups with $|g_{i,1}^K| - 1$ and $|g_{i,2}^K| - 1$ elements, respectively. The product accounts for all possible orderings. □

DEFINITION 5. For any element $g^T \in \mathcal{G}_{n,C}^T$, let $\mathrm{TS}(g^T)$ denote the corresponding (unranked) tree shape $g^{\mathrm{TS}} \in \mathcal{G}_{n,C}^{\mathrm{TS}}$, obtained by removing the rankings from internal nodes of g^T. We call g^T and g'^T equivalent if $\mathrm{TS}(g^T) = \mathrm{TS}(g'^T)$, and we denote the size of the equivalence class by $c^{\mathrm{TS}}(g^T)$.

PROPOSITION 2. *Let $g^T \in \mathcal{G}_{n,C}^T$, and let $g_{i,1}^T$ and $g_{i,2}^T$ be the two subtrees (or clades) that merge at the ith coalescent event, then*

$$c^{\mathrm{TS}}(g^T) = \prod_{i=1}^{n-1} \frac{(|g_{i,1}^T| + |g_{i,2}^T| - 2)!}{(|g_{i,1}^T| - 1)!(|g_{i,2}^T| - 1)!} \left(\frac{1}{2}\right)^{1\{g_{i,1}^T = g_{i,2}^T\}},$$

where $|g_{i,j}^T|$ denotes the number of leaf nodes of $g_{i,j}^T$.

PROOF. Again, the formula is a product of permutations with repetitions. If the two subtrees that merge at the ith coalescence are equal, we need to divide by two since the same rankings in the two subtrees are indistinguishable. □

Given $c^{\mathrm{LT}}(g^K)$ and $c^{\mathrm{TS}}(g^T)$, we can easily compute $q(g^{\mathrm{LT}}) = c^{\mathrm{LT}}(g^K)q(g^K)$ and $q(g^{\mathrm{TS}}) = c^{\mathrm{TS}}(g^T)q(g^T)$. These two distributions constitute our sampling proposal in SIS procedure to estimate $|\mathcal{G}_{n,C}^{\mathrm{LT}}|$ and $|\mathcal{G}_{n,C}^{\mathrm{TS}}|$.

4. Simulations. We rely on simulations to assess the convergence, empirical accuracy and computational performance of the proposed algorithms. We discuss a range of scenarios designed to capture a variety of settings encountered in applications and to highlight the key properties of the algorithms. All the algorithms are implemented in the R package `phylodyn` which is available for download at https://github.com/JuliaPalacios/phylodyn. The four tree topologies analyzed are: $\mathcal{G}_{n,C}^K$: Kingman ranked labeled trees, $\mathcal{G}_{n,C}^T$: Tajima ranked tree shapes, $\mathcal{G}_{n,C}^{\mathrm{TS}}$: tree shapes and $\mathcal{G}_{n,C}^{\mathrm{LT}}$: unranked labeled trees. All of which are compatible with the simulated dataset.

We encode our simulated molecular data as an $n \times m$ incidence matrix $\mathbf{Y}$ of n sequences at m polymorphic sites. $\mathbf{Y}$ is generated in three steps: we first simulate a Kingman genealogy of n individuals; then, we draw m mutations from a Poisson distribution and, finally, we uniformly allocate the m mutations along the branches of the genealogy, that is,

$$\begin{aligned} (g^K, \mathbf{t}) &\sim \text{Kingman } n\text{-coalescent}, \\ m &\sim \text{Poisson}(\mu L), \quad L = \sum_{k=2}^{n} k t_k, \\ x_1, \ldots, x_m &\sim \text{Uniform}(g^K), \end{aligned}$$

where μ denotes the mutation parameter, L the tree length, $\mathbf{t}$ is the $(n-1)$-vector of coalescent times and $x_1, \ldots, x_m$ are the allocations of the mutations along g^K. Coalescent times are exponentially distributed with rate $\binom{k}{2}$ (assuming constant population size). We simulate Kingman genealogies with the open source implementation `R-ape:rcoal()` (Paradis, Claude and Strimmer (2004)). The m mutations are then placed uniformly at random along the branches of the timed genealogy $(g^K, \mathbf{t})$ and labeled $1, \ldots, m$. The matrix $\mathbf{Y}$ is constructed by setting the (i, j)th entry equal to 1 if the branch path from leaf i to the root has labeled

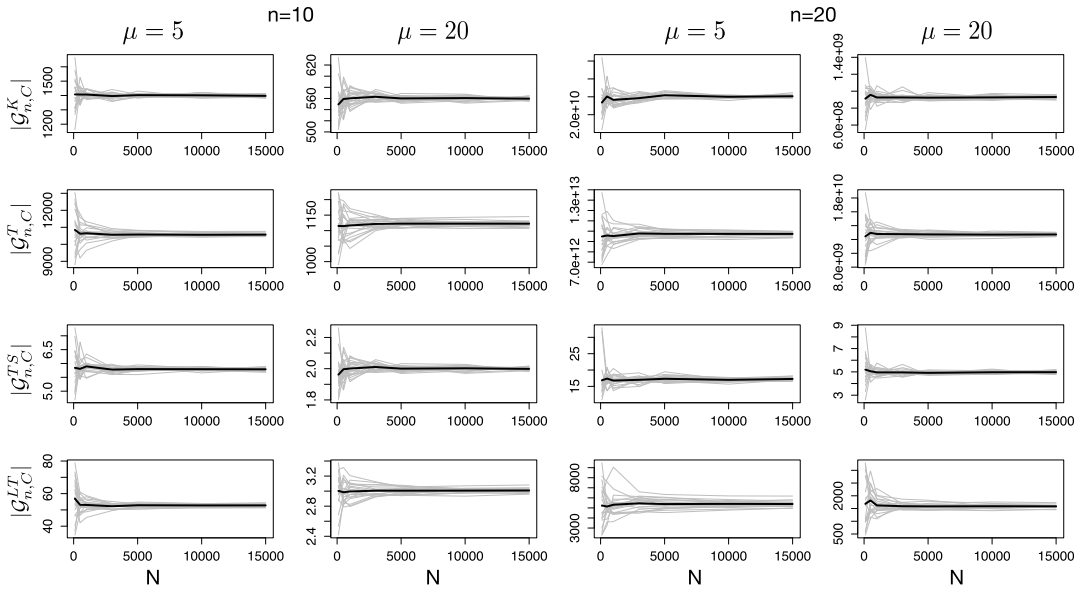

FIG. 8. *Simulations*: *Sequential importance sampling count estimates*. *Rows show the estimated cardinality of the four tree topologies*: *ranked tree shapes* ($|\mathcal{G}^T_{n,C}|$), *ranked labeled trees* ($|\mathcal{G}^K_{n,C}|$), *unranked tree shapes* ($|\mathcal{G}^{\mathrm{TS}}_{n,C}|$) *and labeled trees* ($|\mathcal{G}^{\mathrm{LT}}_{n,C}|$) (*top to bottom rows*); *the first two columns correspond to simulations based on* $n = 10$ *samples and the last two columns on* $n = 20$ *samples*. *The first and third columns correspond to mutation rate* $\mu = 5$ *and second and fourth to* $\mu = 20$. *Grey lines correspond to each of the* 20 *independent estimates from the* 20 *repetitions of the SIS algorithm computed at* $N \in (100, 500, 1000, 3000, 5000, 10{,}000, 15{,}000)$ *iterations*. *Black lines show the mean estimate of the* 20 *repetitions*.

mutation x_j. This part of the simulation algorithm corresponds to the infinite-sites mutation model. Finally, $\mathbf{Y}$ is summarized by its unique set of haplotypes (rows) with corresponding frequencies (incidence matrix in Figure 4). The corresponding perfect phylogeny $\mathcal{T}$ (Figure 4(a)) is constructed via Gusfield (1991) algorithm; the Kingman's perfect phylogeny $\mathcal{T}^K$ and the Tajima's perfect phylogeny $\mathcal{T}^T$ are constructed from $\mathcal{T}$ as described in Section 2.2.

To assess convergence of our algorithms at various sample sizes and with different combinatorial constraints (defined by the patterns of mutations), we simulate incidence matrices under four scenarios, with sample sizes $n \in (10, 20)$ and two mutation regimes $\mu \in (5, 20)$. We computed SIS estimates and diagnostics after N number of iterations with $N \in (100, 500, 1000, 3000, 5000, 10{,}000, 15{,}000)$ from 20 repetitions of each of the four simulation scenarios.

Figure 8 shows the estimated cardinalities (grey lines) of the four topological spaces (rows) and for the four combinations of n and μ (columns). Black lines depict the mean estimates. Figure 9 plots the ratio of the standard error to the estimated counts (relative SE, rSE) averaged over the 20 SIS runs for each coalescent resolution (distinct line types) and the four simulation scenarios (distinct panels). The relevant information in Figure 9 is not the decay of the lines, which is expected, but rather the order of magnitude of the rSE values when comparing the values across the four algorithms (a lower value means a better empirical performance). Table 1 reports the cv^2 values for the four algorithms (rows) and the four combinations of μ and n (columns).

Figure 8 provides a visual inspection of the number of MC samples required for convergence. The four algorithms estimates stabilize around the mean in the four simulation scenarios after $N = 5000$. The variable degrees at which the grey lines are relatively scattered around the mean hints at how the empirical convergence rates deteriorate for larger sample sizes and unranked topologies. Both rSE (Figure 9) and cv^2 (Table 1) confirm and quantify

TABLE 1

Simulations: cv^2 *of estimated cardinalities for the four resolutions. Mean* cv^2 *over* 20 *realizations for* N *in* $\{5000, 10{,}000, 15{,}000\}$. *The four simulated incidence matrices are fixed per each combination of* n *in* $\{10, 20\}$ *and* μ *in* $\{5, 20\}$

	$n = 10$		$n = 20$	
	$\mu = 5$	$\mu = 20$	$\mu = 5$	$\mu = 20$
Tajima algorithm	0.716	0.532	3.674	4.554
Kingman algorithm	0.906	0.698	4.085	3.844
Tree shapes algorithm	1.345	0.532	9.068	5.499
Labeled trees algorithm	4.074	1.096	23.821	26.815

this observation. In Figure 9 we observe an increase of rSE with sample size (two to four times higher when increasing n from 10 to 20). Similarly, the rSEs of unranked counts (dotted and dot-dashed lines) are one to four times higher than the ranked counterparts (solid and dashed lines). Similarly, the cv^2 values in Table 1 increase for larger sample sizes (four to 20 times higher) and for unranked algorithms (one to eight times higher than the cv^2 values for ranked algorithms).

These two trends are expected and have a clear explanation. The effect of a sample size increase is twofold: first, the state space (not conditioning on the data) becomes larger; second, the expected number of mutations increases. A higher number of mutations is likely to impose more constraints in the spaces of genealogies and, consequently, undermine the algorithm performance by having proposals q that "move away" from the uniform distribution (we will elaborate on this point below). The poor performance of both unranked algorithms is due to the fact that we are not sampling from proposal distributions designed to be close to the uniform discrete on the underlying spaces, but, instead, the estimates are obtained by correcting the ranked estimates through the equivalence classes coefficients defined in Propositions 1 and 2 (Section 3.1.3).

A second result is the superior performance of the tree shape algorithm (dotted line in Figure 9 and third row in Table 1) when compared to the labeled tree algorithm (dot-dashed line in Figure 9 and last row in Table 1). The observed result suggests that the tree shape sampling distribution has a better proposal distribution. We hypothesize that this result is a consequence of the fact that the equivalence classes in the unranked spaces are, for most

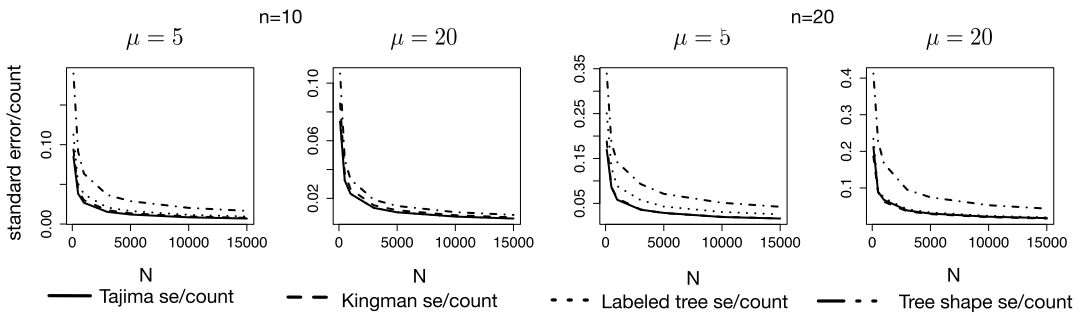

FIG. 9. *Simulations*: *ratio of standard error to approximate count. The lines correspond to the four topologies*: *ranked tree shapes* (*solid*), *ranked labeled trees* (*dashed*), *unranked tree shapes* (*dotted*) *and labeled trees* (*dot-dashed*) *respectively. The first two columns correspond to simulations based on* $n = 10$ *samples and the last two columns on* $n = 20$ *samples. The first and third columns correspond to mutation rate* $\mu = 5$ *and second and fourth correspond to* $\mu = 20$. *Lines plot the ratio of the standard error to the average estimated count over the* 20 *repetitions of the SIS algorithms computed at* $N \in (100, 500, 1000, 3000, 5000, 10{,}000, 15{,}000)$ *iterations. Note that solid and dashed lines practically overlap.*

cases, smaller than in the labeled spaces, that is, $c^{\mathrm{TS}}(g^T) < c^{\mathrm{LT}}(g^K)$, that is, we are sampling in a space that is more similar to the correct one. In contrast, we note that neither cv^2 (Table 1) nor rSE (black and blue lines in Figure 9) provides evidence of different performance between the two ranked algorithms (solid and dashed lines in Figure 9 overlap).

Table 1 highlights a very poor performance of the unranked methods for $n = 20$, especially for labeled trees. A cv^2 close to 20 raises serious concerns on the reliability of the unranked algorithms in this context, suggesting the risk of a variance explosion (cv^2 is a rescaled variance) and low efficiency (ESS is about 5% of the chosen N). Similar cv^2 has been achieved by modern algorithms in real network applications (Chen and Chen (2018)). Lastly, although we do not show plots of Chatterjee and Diaconis (2018)'s q_N and Q_N, these statistics exhibit similar relative performance and identical patterns as those highlighted for rSE and cv^2.

This first simulation study does not assess the variance and scalability of our proposals for different data sets; in particular, it does not assess how the quality of the proposals depends on the perfect phylogeny $\mathcal{T}$. To study this question, we simulate 20 incidence matrices with fixed sample size $n = 15$ and for each μ in $\{1, 4, 7, 10, 13\}$ (100 matrices in total). SIS estimated cardinality and diagnostics are computed at $N = 5000$. In Figure 10(a) we show the cv^2 of the estimated cardinality of ranked tree shapes (Tajima algorithm) as a function of the total number of nodes and the total number of leaf nodes of $\mathcal{T}^T$. We do not plot the cv^2 values of the other three resolutions; however, the other resolutions mirror observed values for the Tajima algorithm. In Figure 10(b) we plot the total computing time (in minutes) to obtain the five statistics: count estimates, cv^2, rSE, ESS and Q_N, for all topologies. In both panels of Figure 10, each dot corresponds to one incidence matrix. The mutation parameter μ is not displayed; it is varied solely to have diverse incidence matrices.

Figure 10 (a) shows a nonlinear relationship between the cv^2 and both the number of nodes and leaf nodes. As expected, when the number of nodes (leaf or total) is low, the cv^2 values are always low. In this case our SIS proposal distributions are close to the Kingman and Tajima jump chains which are uniform and close to uniform distributions, respectively, on the space of trees. As the number of nodes increases, the cv^2 values exhibit a large variation across data sets. In this case the quality of our proposal deteriorates for a few datasets substantially. Our

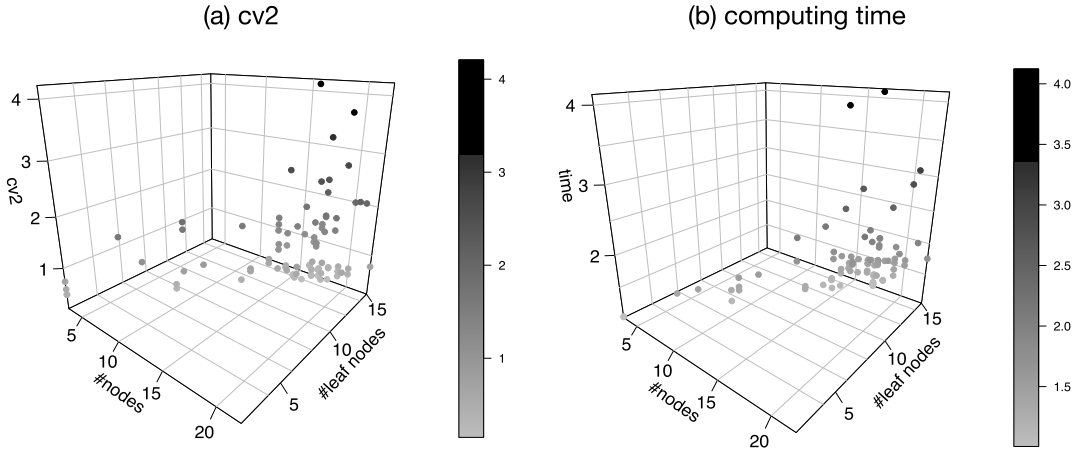

FIG. 10. *Simulations*: cv^2 *of the Tajima algorithm* (*a*) *and total* (*all topological spaces*) *computing time* (*b*). *Panel* (*a*) *plots the* cv^2 *of that Tajima algorithm as a function of the number of nodes* (# *nodes*) *and the number of leaf nodes* (# *leaf nodes*) *in the perfect phylogeny* $\mathcal{T}^T$. *Panel* (*b*) *reports the computing time to obtain the estimate counts for the four topologies and the convergence diagnostics* (*rSE*, *ESS*, q_N, cv^2). *Each dot corresponds to an algorithm run* ($N = 5000$) *for each of the* 100 *incidence matrices simulated with fixed sample size* $n = 15$ *and* μ *in* $\{1, 4, 7, 10, 13\}$. *Dots with different gray scale intensities represent the numerical value of* cv^2 (*Panel* (*a*)) *and total time in minutes* (*Panel* (*b*)) *as represented by the vertical bars to the left of the plots*.

interpretation is that more mutation constraints move our proposal distributions further away from the uniform distribution.

A second result of the simulation study is that the Kingman algorithm has a better performance, on average, than the Tajima algorithm. The mean of the cv^2 values across different data sets is 1.02 for the Tajima algorithm and 0.72 for the Kingman algorithm. This is coherent with the construction of our SIS proposals and the fact that the Kingman jump chain is exactly uniform.

Lastly, the computing time (Figure 10(b)) exhibits exactly the same observed patterns as cv^2 (Figure 10(a)). Longer computing times are driven mostly by the backtracking algorithm. Trivially, the more nodes in $\mathcal{T}^T$, the more nodes ordering $\mathbf{v}$ the backtracking algorithm may need to explore. However, we note that the computing time remains low for most data sets, even for data sets with large number of nodes.

5. Case studies.

5.1. *Case study* 1: *Multiresolution simulation study.* As discussed in the Introduction, there is a growing interest in population genetics to use more efficient lower resolution coalescent models for inferring evolutionary parameters from molecular data (Palacios et al. (2019+), Sainudiin, Stadler and Véber (2015), Sainudiin and Véber (2018)). However, no work has been done to quantify the "real" gains of working with different coalescent resolutions to real data. It is important to address this question before this research direction is further explored. This case study addresses this question through simulations.

Data. We simulate 50 incidence matrices for each of 24 possible pairings of n in $(5, 10, 15, 20)$ and μ in $(2, 5, 10, 20, 50, 75)$. For each simulated dataset we estimate the cardinality of the four constrained topological spaces. Based on the results observed in the previous section, we set $N = 5000$ for $n \in (5, 10)$, $N = 10{,}000$ for $n = 15$, and $N = 15{,}000$ for $n = 20$.

Results. In the first row of Figure 11, we show the log ratio of the estimated cardinalities of Kingman topologies to Tajima topologies, and, in the second row, we show the log ratio of the estimated cardinalities of labeled trees to tree shapes. Table 2 summarizes the results for a single iteration picked at random from the 50 replicates. We note that an average over the 50 replicates is not insightful given the high variability of the incidence matrices sampled (which can be observed by the length of the boxplots in Figure 11).

Figure 11 shows that the cardinality of the space of Kingman trees is always larger than the cardinality of the space of Tajima trees (first row), and the cardinality of the space of labeled trees is always larger than the cardinality of the space of tree shapes (second row). As mentioned in the Introduction, this is relevant for population genetic studies that aim to estimate evolutionary parameters by integrating over the space of trees. When analyzing how much effective reduction in the tree space is gained by assuming the infinite sites model alone, we note that a high mutation rate will, in general, constrain the tree sample space more than a low mutation rate. This reduction is accentuated for Kingman's trees under every simulation scenario. For example, for $n = 20$ there are 5.64×10^{29} (exact) unconstrained Kingman's trees (using formula from Section 2). This number drops to $5.67 \times 10^{10} \pm 2.08 \times 10^{9}$ (SIS estimate) for a simulated dataset with $\mu = 20$. The (exact) unconstrained number of ranked tree shapes is 2.9×10^{13} which drops to $4.63 \times 10^{10} \pm 1.47 \times 10^{8}$ (SIS estimate) for a simulated dataset with $\mu = 20$. A similar pattern is observed for unranked tree shapes.

For a fixed sample size, we observe that the difference between Kingman and Tajima cardinalities decays exponentially from low mutation regimes to high mutation regimes (moving along x axes in the plots of the first row in Figure 11). The same trend is observed between labeled topologies and tree shapes (moving along x axes in the plots of the second row of

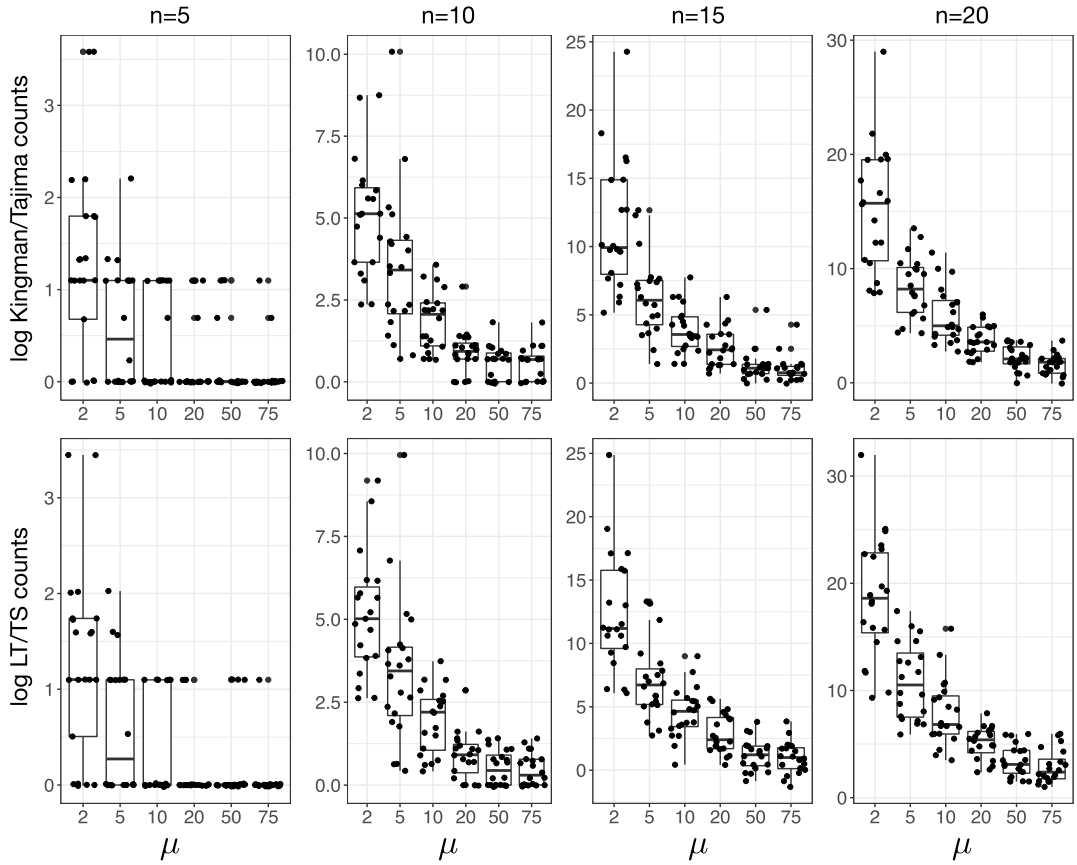

FIG. 11. *Case study* 1. *Multiresolution simulation study*: *Log ratio of estimated counts for varying n and* μ. *Rows correspond to the log ratio of cardinalities between Kingman and Tajima topologies* (*first row*) *and the log ratio of cardinalities between labeled trees and tree shapes* (*second row*). *Columns represent different sample sizes n and boxplots within each plot show results under different mutation rates. Boxplots are generated from* 50 *independent simulations. Dots represent the SIS count estimates computed for* $N = 5000$ (*for* $n = 5, 10$), $N = 10{,}000$ (*for* $n = 15$), *and* $N = 15{,}000$ (*for* $n = 20$). *Dots are spread over the box width for ease of visualization.*

Figure 11). For example, keeping $n = 20$ fixed, the Tajima space is on average approximately 2×10^{11} smaller than the Kingman space for $\mu = 2$, 6900 times smaller for $\mu = 10$, and only *seven* times smaller for $\mu = 75$.

For a fixed mutation regime, the respective differences between labeled topologies counts and unlabeled topologies counts (Kingman vs. Tajima, labeled trees vs tree shapes) becomes more pronounced as we increase the sample size (columns in Figure 11.) For example, keeping $\mu = 10$ fixed, the Tajima space is on average approximately 1.7 times smaller than the Kingman space for $n = 5$, 225 times smaller for $n = 10$ and only 6900 times smaller for $n = 20$.

The case study suggests that modeling with lower resolution coalescent models (unlabeled) could be advantageous when applied to organisms with low mutation rates such as humans or mammals. However, the advantages are less pronounced for rapidly evolving organisms such as pathogens and viruses.

5.2. *Case study* 2: *Human mitochondrial and nuclear DNA data*. Present day molecular data at a nonrecombining segment (or locus) from a sample of individuals inform about past population history and other evolutionary parameters (Tavaré (2004)). Multiple independent loci, perhaps loci at different chromosomes or loci from distant locations across the genome,

TABLE 2

Case study 1. *Multiresolution simulation study*: *SIS counts for varying sample size* (n) *and mutation rate* (μ). n *denotes sample size*, μ *mutations rate*, $|J_{\text{leaf}}|$ *denotes the number of leaf nodes in* $\mathcal{T}$, $|J|$ *denotes the number of nodes in* $\mathcal{T}$. *Counts are reported for the four resolutions plus/minus the standard error*

n	μ	$\lvert J\rvert$	$\lvert J_{\text{leaf}}\rvert$	Tajima trees	Kingman trees	Labeled trees	Tree shapes
5	2	7	5	4.979 ± 0.02536	30.05 ± 0.2057	14.97 ± 0.1832	2.629 ± 0.02622
	5	8	5	2.993 ± 0.005794	8.988 ± 0.01735	2.996 ± 0.005785	0.9976 ± 0.001931
	10	8	5	2.999 ± 0.005778	8.99 ± 0.01735	2.997 ± 0.005784	0.9996 ± 0.001926
	20	9	5	3.01 ± 0.01414	3.024 ± 0.01414	1.008 ± 0.004713	1.003 ± 0.004714
	50	9	5	3.016 ± 0.01414	3.025 ± 0.01414	1.008 ± 0.004713	1.005 ± 0.004714
	75	9	5	3.018 ± 0.01414	2.976 ± 0.01414	0.9919 ± 0.004713	1.006 ± 0.004714
10	2	14	9	499.9 ± 7.26	5350 ± 91.37	21.72 ± 0.6739	1.565 ± 0.03521
	5	14	9	509.3 ± 7.354	5367 ± 91.69	21.94 ± 0.6798	1.561 ± 0.03405
	10	15	9	499.7 ± 8.79	1765 ± 35.78	7.134 ± 0.2438	1.471 ± 0.03291
	20	16	9	430.7 ± 5.871	1235 ± 17.07	2.94 ± 0.04064	1.026 ± 0.01398
	50	16	9	422.8 ± 5.822	1235 ± 17.07	2.94 ± 0.04064	1.007 ± 0.01386
	75	18	10	418 ± 5.724	1249 ± 17.35	2.974 ± 0.04131	0.9952 ± 0.01363
15	2	15	11	$3{,}474{,}000 \pm 53{,}560$	$1.112\text{e}{+10} \pm 141{,}400{,}000$	$1{,}087{,}000 \pm 46{,}170$	65.74 ± 1.95
	5	20	12	$297{,}200 \pm 6108$	$3{,}318{,}000 \pm 68{,}620$	603.1 ± 36.96	3.962 ± 0.1131
	10	21	12	$60{,}630 \pm 1475$	$650{,}800 \pm 14{,}850$	434.2 ± 13.6	1.902 ± 0.05244
	20	22	12	$60{,}330 \pm 1386$	$226{,}000 \pm 5297$	147.4 ± 4.842	1.838 ± 0.04675
	50	24	13	$45{,}240 \pm 961.4$	$141{,}600 \pm 3048$	45.3 ± 0.9987	1.004 ± 0.02134
	75	26	14	$43{,}410 \pm 894.5$	$141{,}100 \pm 3036$	45.1 ± 0.9922	0.9637 ± 0.01986
20	2	20	15	$8.05\text{e}{+11} \pm 1.571\text{e}{+10}$	$1.731\text{e}{+17} \pm 2.699\text{e}{+15}$	$1.588\text{e}{+10} \pm 4.211\text{e}{+09}$	2084 ± 86.73
	5	23	14	$8.429\text{e}{+09} \pm 209{,}800{,}000$	$6.869\text{e}{+11} \pm 1.56\text{e}{+10}$	32250 ± 1846	16.58 ± 0.6664
	10	23	13	$4.339\text{e}{+09} \pm 83{,}760{,}000$	$1.189\text{e}{+11} \pm 2.274\text{e}{+09}$	5485 ± 300.7	5.19 ± 0.1093
	20	28	16	$4.344\text{e}{+09} \pm 95{,}910{,}000$	$2.757\text{e}{+10} \pm 621{,}200{,}000$	1237 ± 85.72	5.102 ± 0.1238
	50	28	15	$437{,}600{,}000 \pm 9{,}594{,}000$	$2.816\text{e}{+09} \pm 6{,}4850{,}000$	335.4 ± 9.758	0.9448 ± 0.02071
	75	32	17	$458{,}700{,}000 \pm 10{,}720{,}000$	$2.754\text{e}{+09} \pm 60{,}920{,}000$	335.7 ± 9.126	0.9904 ± 0.02315

provide multiple independent realizations of the same coalescent process with shared population history. In this case study we show that, under the infinite sites model, independent chromosomal regions can impose a completely different set of constraints on their local tree topology. We provide quantitative evidence of this effect. We note that these constraints do not arise employing alternative mutation models, for example, Jukes–Cantor. We apply our method to mitochondrial DNA (mtDNA) and nuclear DNA (nDNA) data. mtDNA is known to have a much higher mutation rate than nDNA (Song et al. (2005)), and, thus, we expect a larger reduction in the state space. In addition, we explore how these constraints vary across resolutions.

Data. We analyze $n = 30$ samples of mitochondrial DNA (mtDNA) selected uniformly at random from the 107 Yoruban individuals available in the 1000 Genomes Project phase 3 (1000 Genomes Project Consortium (2015)). We retained the coding region: 576–16,024 according to the rCRS reference of Human Mitochondrial DNA (Anderson et al. (1981), Andrews et al. (1999)) and removed 38 indels (insertions and deletions are not modeled in our approach). Of the 260 polymorphic sites, we only retained 240 sites compatible with the infinite sites mutation model. Ancestral states (0s in the incidence matrix) were obtained from the RSRS root sequence (Behar et al. (2012)). For nDNA we analyze 2320 sites of the β-globin gene in chromosome 11 from $n = 30$ Melanesian individuals subsampled from a larger incidence matrix ($n = 57$) analyzed in Griffiths and Tavaré (1999). It was already part of a larger dataset described in Harding et al. (1997). Figure 12 plots the Tajima perfect phylogeny for the two datasets. The mtDNA comprises 29 haplotypes, and the nDNA comprises *four* haplotypes.

Results. We estimated the cardinalities of the four constrained topologies at $N = 35{,}000$. The number of iterations N is chosen by the criteria discussed in the simulation section.

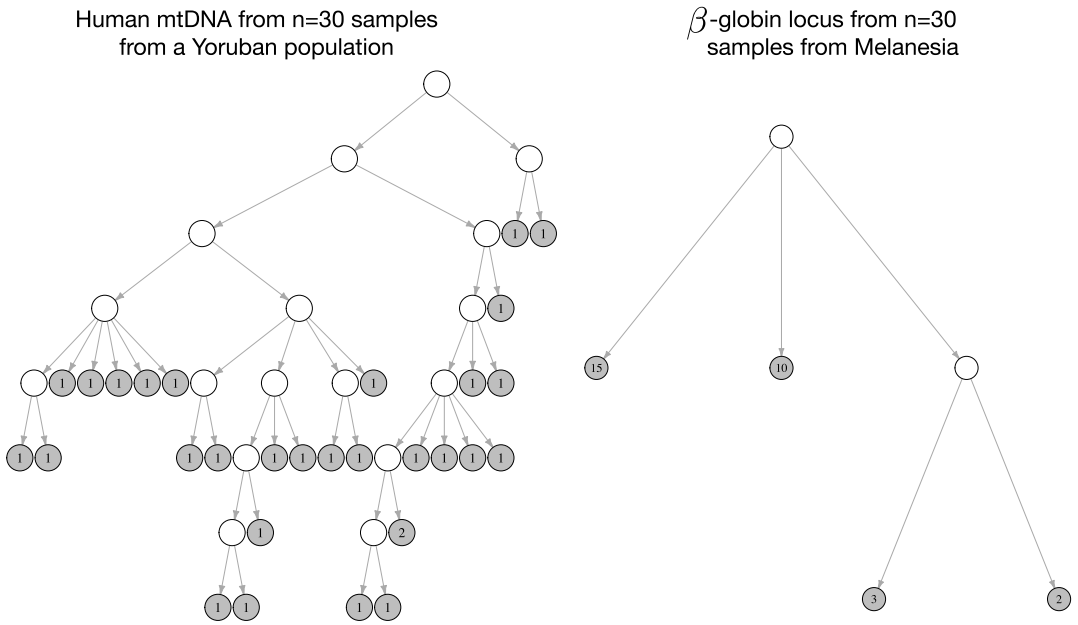

FIG. 12. *Case study* 2: *Tajima perfect phylogenies of Yoruban mitochondrial data* (*left*) *and Melanesian* β*-globin locus data* (*right*). *Left panel*: $\mathcal{T}^T$ *of* $n = 30$ *sequences of mtDNA sampled at random from* 107 *Yoruban individuals available from the* 1000 *Genomes Project phase* 3 (1000 *Genomes Project Consortium* (2015)). *Right panel*: $\mathcal{T}^T$ *of* $n = 30$ *sequences of DNA from the* β*-globin locus sampled at random from* 57 *Melanesian individuals available in Fullerton et al.* (1994). *Grey nodes represent the leaf nodes. The number within a node is the number of individuals assigned to that node.*

TABLE 3

Case Study 2: *Estimated counts and* cv^2 *for the mtDNA and nDNA datasets. Unconstrained refers to the size of the underlying tree space* (*rows*) *cardinalities when we are not conditioning on the data. Estimates are obtained with* $N = 35{,}000$; $\pm$ *adds or subtracts the standard error*

		Yoruban mtDNA		β-globin locus nDNA	
Dataset ($n = 30$)	Unconstrained	Estimate	cv^2	Estimate	cv^2
Tajima	2.31×10^{25}	$1.05 \times 10^{20} \pm 6.19 \times 10^{18}$	69.2	$3.10 \times 10^{23} \pm 2.21 \times 10^{21}$	1.78
Kingman	4.37×10^{54}	$7.17 \times 10^{23} \pm 3.01 \times 10^{22}$	36.9	$1.07 \times 10^{40} \pm 4.68 \times 10^{37}$	0.66
Tree shapes	1.41×10^{9}	$1.33 \times 10^{3} \pm 1.18 \times 10^{2}$	165.9	$3.10 \times 10^{6} \pm 2.81 \times 10^{5}$	343.6
Labeled trees	4.95×10^{38}	$1.17 \times 10^{12} \pm 2.10 \times 10^{11}$	674.1	$4.65 \times 10^{27} \pm 1.49 \times 10^{27}$	372.6

Table 3 shows the size of the unconstrained spaces, along with our SIS estimates and cv^2 values.

First, the spaces of trees compatible with the β-globin dataset are many orders of magnitude larger than the spaces of trees compatible with the mtDNA dataset. Following Case Study 1, this effect could have been predicted by the lower number of segregating sites. Second, results in Table 3 gives a different perspective on the computational limits of coalescent based inference: under Kingman coalescent (still the dominant model in the field of population genetics), the sample space of trees for the mtDNA is massively smaller than the unconstrained space; it drops from 4.37×10^{54} to $7.17 \times 10^{23} \pm 3.01 \times 10^{22}$. Whereas, when the presence of a reduction was known, such a reduction had never been quantified.

With respect to the performance of our algorithms, we confirm that the variance of the Kingman and the Tajima algorithms (ranked tree topologies) are mostly determined by the perfect phylogeny structure rather than the sample size: the cv^2 for mtDNA is larger than the cv^2 for nDNA (first and second rows of Table 3); in particular, the nDNA data (four leaf nodes) exhibits very low cv^2 (second row in Table 3). The large cv^2 values obtained with the unranked algorithms (third and fourth rows) questions the validity of our estimated counts. We note that the order of magnitude of the estimate is more meaningful than the point estimate itself; the reductions in cardinality with respect to the unconstrained size are all consistent with theoretical expectations and the simulation studies. Similarly, the reductions in cardinalities across resolutions are more extreme in the mtDNA dataset than in the β-globin dataset. Surely, this case study displays a situation where a sequential importance sampler experiences variance explosion.

6. Discussion. In this article we propose a set of algorithms to sequentially sample tree topologies compatible with the observed data. We use our sampling algorithms to estimate the cardinality of the sample space of tree topologies with importance sampling. We assume that our sampled locus is nonrecombining and that the infinite sites assumption holds. In the infinite sites mutation model, each site in the locus can mutate only once. While in practice it is possible to observe sites that are not compatible with this mutation model, the percentage of these cases is usually marginal for some organisms, such as humans and other primates. The major implication of the infinite sites mutation model is that observed data impose constraints on the space of compatible trees. We analyze the cardinality of the following constrained tree spaces: ranked labeled trees (Kingman), ranked tree shapes (Tajima), unranked labeled trees and tree shapes. These sample tree spaces correspond to different resolutions of the n-coalescent process.

Our proposed algorithms sample a tree topology in a bottom-up fashion: given a sample of n individuals, we sequentially build the trees in $n - 1$ steps. We employ a graphical representation of the data called perfect phylogeny that allows us to account for the combinatorial

constraints imposed by the data. The perfect phylogeny "groups" individuals in different nodes; in our algorithms coalescent events are allowed solely among individuals assigned to the same node. Within each node the choice of which individuals coalesce is regulated by the underlying jump chain of the coalescent process we are modeling.

The research question tackled in this paper was motivated by the challenging inference problem of coalescent methods used in population genetics. There is a growing interest in exploring different resolutions of the n-coalescent process for inference of evolutionary parameters from molecular sequence data in order to gain computational tractability. Indeed, the size of the hidden state-space of trees in the standard Kingman coalescent grows superexponentially with the sample size. Despite the a priori reduction in the cardinality of the state space obtained by using coarser modeling resolutions, for example, Tajima n-coalescent, a quantification of this reduction conditionally on the data was unknown. Given the amount of work and software available tailored to the Kingman n-coalescent, it was in our opinion fundamental to quantify the benefits of modeling with different resolutions before any more work is carried out.

From our empirical analyses, it emerges that the benefits of using a coarser resolution depend largely on the data considered. The advantages are striking as the sample size increases, especially in regimes of low mutation rate such as in nuclear human DNA variation. In general, the greater the number of observed mutations is, the less are the benefits of employing coarser resolutions. This is consistent with theoretical predictions, under the infinite sites assumption, mutations induce some labeling and individuals can be distinguished according to private mutations. In this case the benefits of employing an unlabeled tree are less evident. This observation applies to both ranked and unranked trees. In applications where the number of mutations is low, the benefits of coarser resolutions remain clear.

In the context of recombination, the perfect phylogeny is no longer a single tree but a set of trees called *perfect phylogeny forest* (Gusfield (2014)), and we believe that our methodology can be extended in this context. Indeed, many new interesting research questions open up; for example, the number of trees in the forest, that is, the number of recombination events which is itself a challenging problem known as the *minimum perfect phylogenetic forest problem*. In this case the target is not a space of tree genealogy but a space of networks known as *the ancestral recombination graph* (Griffiths and Marjoram (1997)) and a future area of research.

Acknowledgments. We would like to acknowledge Persi Diaconis who brought our attention to the use of sequential importance sampling for approximate counting. This work is supported by R01 GM131404 and the Alfred P. Sloan Foundation. We would like to acknowledge two anonymous reviewers for their suggestions that greatly improved the manuscript.

REFERENCES

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526** 68 EP.

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

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23** 147. https://doi.org/10.1038/13779

Behar, D. M., Van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N. M., Kivisild, T., Torroni, A. and Villems, R. (2012). A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90** 675–684.

Blanchet, J. and Rudoy, D. (2009). Rare event simulation and counting problems. In *Rare Event Simulation Using Monte Carlo Methods* 171–192. Wiley, Chichester. MR2730766 https://doi.org/10.1002/9780470745403.ch8

BLITZSTEIN, J. and DIACONIS, P. (2010). A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Math.* **6** 489–522. MR2809836 https://doi.org/10.1080/15427951.2010.557277

CAYLEY, A. (1856). Note sur une formule pour la reversion des séries. *J. Reine Angew. Math.* **52** 276–284. MR1578984 https://doi.org/10.1515/crll.1856.52.276

CHATTERJEE, S. and DIACONIS, P. (2018). The sample size required in importance sampling. *Ann. Appl. Probab.* **28** 1099–1135. MR3784496 https://doi.org/10.1214/17-AAP1326

CHEN, Y. and CHEN, Y. (2018). An efficient sampling algorithm for network motif detection. *J. Comput. Graph. Statist.* **27** 503–515. MR3863753 https://doi.org/10.1080/10618600.2017.1391696

CHEN, Y., DIACONIS, P., HOLMES, S. P. and LIU, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *J. Amer. Statist. Assoc.* **100** 109–120. MR2156822 https://doi.org/10.1198/016214504000001303

DIACONIS, P. (2018). Sequential importance sampling for estimating the number of perfect matchings in bipartite graphs: An ongoing conversation with Laci. Preprint.

DISANTO, F. and WIEHE, T. (2013). Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. *Math. Biosci.* **242** 195–200. MR3068684 https://doi.org/10.1016/j.mbs.2013.01.010

DRUMMOND, A. J., SUCHARD, M. A., XIE, D. and RAMBAUT, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29** 1969–1973.

FERRETTI, L., LEDDA, A., WIEHE, T., ACHAZ, G. and RAMOS-ONSINS, S. E. (2017). Decomposing the site frequency spectrum: The impact of tree topology on neutrality tests. *Genetics* **207** 229–240. https://doi.org/10.1534/genetics.116.188763

FULLERTON, S. M., HARDING, R. M., BOYCE, A. J. and CLEGG, J. B. (1994). Molecular and population genetic analysis of allelic sequence diversity at the human beta-globin locus. *Proc. Natl. Acad. Sci. USA* **91** 1805–1809.

GAO, F. and KEINAN, A. (2016). Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. *Genetics* **202** 235–245.

GATTEPAILLE, L., GÜNTHER, T. and JAKOBSSON, M. (2016). Inferring past effective population size from distributions of coalescent times. *Genetics* **204** 1191–1206. https://doi.org/10.1534/genetics.115.185058

GRIFFITHS, R. C. (1987). Counting genealogical trees. *J. Math. Biol.* **25** 423–431. MR0908383 https://doi.org/10.1007/BF00277166

GRIFFITHS, R. C. and MARJORAM, P. (1997). An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution* (*Minneapolis*, *MN*, 1994). *IMA Vol. Math. Appl.* **87** 257–270. Springer, New York. MR1493031 https://doi.org/10.1007/978-1-4757-2609-1_16

GRIFFITHS, R. C. and TAVARÉ, S. (1999). The ages of mutations in gene trees. *Ann. Appl. Probab.* **9** 567–590. MR1722273 https://doi.org/10.1214/aoap/1029962804

GUSFIELD, D. (1991). Efficient algorithms for inferring evolutionary trees. *Networks* **21** 19–28. MR1083125 https://doi.org/10.1002/net.3230210104

GUSFIELD, D. (2014). *ReCombinatorics*: *The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*. MIT Press, Cambridge, MA. With contributions from Charles H. Langley, Yun S. Song and Yufeng Wu. MR3237544

HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1965). *Monte Carlo Methods*. Methuen & Co., Ltd., London; Barnes & Noble, Inc., New York. MR0223065

HARDING, R. M., FULLERTON, S. M., GRIFFITHS, R. C., BOND, J., COX, M. J., SCHNEIDER, J. A., MOULIN, D. S. and CLEGG, J. B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60** 772–789.

JERRUM, M. and SINCLAIR, A. (1996). The Markov chain Monte Carlo method: An approach to approximate counting and integration. In *Approximation Algorithms for NP-Hard Problems* 482–520.

JERRUM, M. R., VALIANT, L. G. and VAZIRANI, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** 169–188. MR0855970 https://doi.org/10.1016/0304-3975(86)90174-X

KIMURA, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61** 893.

KINGMAN, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probab.* **Special Vol. 19A** 27–43. Essays in statistical science. MR0633178

KNUTH, D. E. (1976). Mathematics and computer science: Coping with finiteness. *Science* **194** 1235–1242. MR0534161 https://doi.org/10.1126/science.194.4271.1235

KNUTH, D. E. (2018). *Art of Computer Programming*, *Volume* 4*B*, *Fascicle* 5: *The*: *Mathematical Preliminaries Redux*; *Backtracking*; *Dancing Links*. Addison Wesley Professional.

LIU, D., SHI, W., SHI, Y., WANG, D., XIAO, H., LI, W., BI, Y., WU, Y., LI, X. et al. (2013). Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: Phylogenetic, structural, and coalescent analyses. *Lancet* **381** 1926–1932.

MALIET, O., GASCUEL, F. and LAMBERT, A. (2018). Ranked tree shapes, nonrandom extinctions, and the loss of phylogenetic diversity. *Syst. Biol.* **67** 1025–1040. https://doi.org/10.1093/sysbio/syy030

NORDBORG, M. (1998). On the probability of Neanderthal ancestry. *Am. J. Hum. Genet.* **63** 1237–1240. https://doi.org/10.1086/302052

OWEN, A. B. (2013). Monte Carlo theory, methods and examples. Online.

PALACIOS, J. A. and MININ, V. N. (2013). Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics* **69** 8–18. MR3058047 https://doi.org/10.1111/biom.12003

PALACIOS, J. A., VÉBER, A., CAPPELLO, L., WANG, Z., WAKELEY, J. and RAMACHANDRAN, S. (2019). Bayesian estimation of population size changes by sampling Tajima's trees. *Genetics* **2**13 967–986.

PARADIS, E., CLAUDE, J. and STRIMMER, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20** 289–290.

ROSENBERG, N. A. and NORDBORG, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3** 380–390. https://doi.org/10.1038/nrg795

SAINUDIIN, R., STADLER, T. and VÉBER, A. (2015). Finding the best resolution for the Kingman–Tajima coalescent: Theory and applications. *J. Math. Biol.* **70** 1207–1247. MR3323594 https://doi.org/10.1007/s00285-014-0796-5

SAINUDIIN, R. and VÉBER, A. (2018). Full likelihood inference from the site frequency spectrum based on the optimal tree resolution. *Theor. Popul. Biol.* **124** 1–40.

SINCLAIR, A. (2012). *Algorithms for Random Generation and Counting*: *A Markov Chain Approach*. Springer Science & Business Media.

SONG, S., PURSELL, Z. F., COPELAND, W. C., LONGLEY, M. J., KUNKEL, T. A. and MATHEWS, C. K. (2005). DNA precursor asymmetries in mammalian tissue mitochondria and possible contribution to mutagenesis through reduced replication fidelity. *Proc. Natl. Acad. Sci. USA* **102** 4990–4995.

STEEL, M. (2016). *Phylogeny—Discrete and Random Processes in Evolution. CBMS-NSF Regional Conference Series in Applied Mathematics* **89**. SIAM, Philadelphia, PA. MR3601108 https://doi.org/10.1137/1.9781611974485.ch1

TAJIMA, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105** 437–460.

TAVARÉ, S. (2004). *Ancestral Inference in Population Genetics. Lectures on Probability Theory and Statistics*: *Ecole D'Eté de Probabilités de Saint-Flour XXXI*-2001. Springer.

TERHORST, J., KAMM, J. A. and SONG, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49** 303–309.

WANG, L., BOUCHARD-CÔTÉ, A. and DOUCET, A. (2015). Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. *J. Amer. Statist. Assoc.* **110** 1362–1374. MR3449032 https://doi.org/10.1080/01621459.2015.1054487

WATTERSON, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7** 256–276. MR0366430 https://doi.org/10.1016/0040-5809(75)90020-9