Sequential importance sampling for multi-resolution Kingman-Tajima coalescent counting

Lorenzo Cappello* and Julia A. Palacios*

*Stanford University

February 15, 2019

Abstract

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 space of genealogical trees under different coalescent resolutions. Our sampling scheme proceeds sequentially across the set of combinatorial constraints imposed by the data. We analyse the cardinality of different genealogical tree spaces on simulations to study the settings that favor coarser resolutions. We estimate the cardinality of genealogical tree spaces from mtDNA data from the 1000 genomes and a sample from a Melanesian population to illustrate the settings in which it is advantageous to employ coarser resolutions.

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 (Nordborg 1998, Rosenberg and Nordborg 2002, Liu et al. 2013). 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 $\mathbf{Y}$ in 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:

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

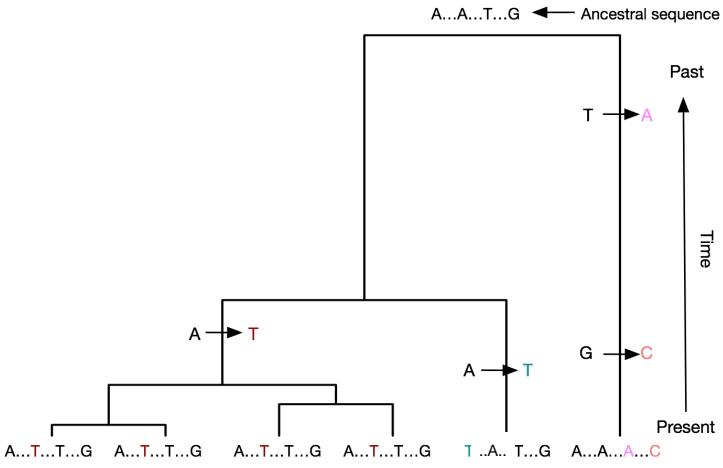

Figure 1: **Coalescence and mutation.** A genealogy of 6 individuals at a locus of 100 base pairs is depicted as a bifurcating tree. Four mutations (at different sites) are superimposed along the branches of the tree giving rise to the 6 sequences shown at the tips of the tree. The 96 sites (base pairs) that do not mutate are represented by dots and only the nucleotides at the polymorphic sites are shown.

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) is usually approximated via Monte Carlo or Markov chain Monte Carlo. 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.

In order to gain computational tractability, researchers have developed both methods that rely on a reduced space of tree topologies, and inferential algorithms (exact or approximate) beyond MCMC. For example, several methods have been proposed to infer $N(t)$ from summary statistics such as the site frequency spectra (Terhorst et al. 2017), from an estimated genealogy (Palacios and Minin 2013, Gattepaille et al. 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 et al. 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 any given dataset and mutation model, the true reduction in cardinality, that is, the number of tree topologies for which $P(\mathbf{Y} \mid \mathbf{g}, \mu) > 0$ (compatible), is not 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 et al. 2015). Reliable estimation of the cardinality of the coalescent hidden state space for a particular data set should provide a valuable guidance to statisticians in designing methods that balance computational efficiency and data sufficiency for inferring evolutionary parameters from different summary statistics. In addition, the cardinality of the topological tree space can also be used directly in inferential algorithms beyond MCMC, such as sequential Monte Carlo methods (Wang et al. 2015).

In this work, the combinatorial question of counting the number of compatible tree topologies with the data is treated as a a statistical problem: estimation of the normalizing constant of a uniform discrete distribution over the space of compatible tree topologies. Our estimation method is an instance of sequential importance sampling (SIS) applied to count discrete structures subject to constraints (Knuth 1976, Chen et al. 2005, Blitzstein and Diaconis 2011, Chen and Chen 2018, Diaconis 2018). More specifically, our algo-

rithm 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:

$$\mathrm{E}_q\left[\frac{1}{q(g)}\right] = \sum_{g\in\mathcal{G}_C} \frac{1}{q(g)} q(g) = |\mathcal{G}_C|, \tag{2}$$

where $\mathcal{G}_C$ is the space of compatible tree topologies.

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, from a sample of human mtDNA from the 1000 genomes and from other human DNA datasets. Section 5 concludes.

2 Preliminaries

2.1 Kingman-Tajima coalescent

Kingman's coalescent $(C(t))_{t\geq 0}$ 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 $\{\{1\}, \ldots, \{n\}\}$ at time $t_n = 0$ (present time at the tips of the tree). As time increases and we go further into the past, the process remains constant until t_{n-1} when two of the n individuals coalescence (represented as the merger of two branches in a single internal node in the genealogy). The state of the process after the first transition (at time t_{n-1}) 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 at t_1 when all individuals coalesce, i.e. at state $\{1, \ldots, n\}$ when there is a single set (at the root of the genealogy when all individuals have a common ancestor).

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, *i.e.* a complete realization of the embedded jump chain of the process $g^K = \{c_j\}_{j=1}^n$. 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:

$$P(\mathcal{C}_{i-1} = c_{i-1} \mid \mathcal{C}_i = c_i) = \begin{cases} \binom{i}{2}^{-1} & \text{if } c_{i-1} \prec c_i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

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 (3) that $P(g^K) = 2^{n-1}/[n!(n-1)!]$, i.e. 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 $(\boldsymbol{\alpha}(t))_{t>0} = (\alpha_1(t), \alpha_2(t))_{t>0}$ is a continuous-time Markov chain that keeps track the number of singletons $\alpha_1(t)$ and the set of extant vintage labels $\alpha_2(t)$ at time t (Tajima 1983, Sainudiin et al. 2015). We refer to singleton as an individual who has not yet coalesced, and a vintage as the internal node of a genealogy labeled by the jump chain step. Since singletons' labels are ignored, there are up to three possible transitions: two singletons merge, one singleton and a vintage merge, or two vintages merge. Formally, given a current state $\alpha_1(t_j)$ and $\alpha_2(t_j)$, when there are $j = \alpha_1(t_j) + |\alpha_2(t_j)|$ branches in

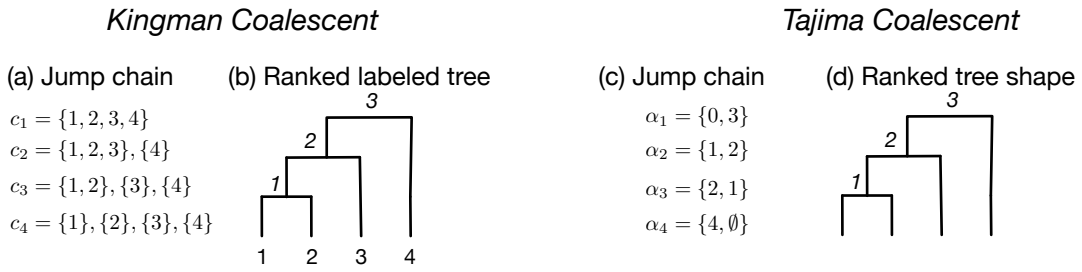

Figure 2: **Coalescent tree topologies. (a)** A complete realization from Kingman's jump chain, and **(b)** its corresponding bijection: a ranked labeled tree topology. **(c)** A complete realization from Tajima's jump chain, and **(d)** its corresponding bijection: a ranked tree shape.

the genealogy, the chain transitions to $\alpha_1(t_{j-1}) = \alpha_1(t_j) - 2$ and $\alpha_1(t_{j-1}) = \alpha_2(t_j) \cup \{j\}$ if two singletons create a new vintage with label $\{j\}$; the chain transitions to $\alpha_1(t_{j-1}) = \alpha_1(t_{j-1}) - 1$ and $\alpha_1(t_{j-1}) = \alpha_2(t_j) \setminus \{i\} \cup \{j\}$ if one singleton and vintage with label $\{i\}$ merge to create a new vintage with label $\{j\}$; and the chain transitions to $\alpha_1(t_{j-1}) = \alpha_1(t_j)$ and $\alpha_1(t_{j-1}) = \alpha_2(t_j) \setminus \{i,k\} \cup \{j\}$ if vintages $\{i\}$ and $\{k\}$ merge to create a new vintage with label $\{j\}$. The process starts at $\alpha_1(0) = n$ and $\alpha_2(0) = \emptyset$ at time $t_n = 0$ (present time at the tips of the tree). As time increases and we go further into the past, the process remains constant until t_n when two singletons coalesce to form a new vintage with label 1. The state of the process after the first transition (at time t_{n-1}) is $\alpha_1(t_{n-1}) = n-2$ and $\alpha_2(t_{n-1}) = \{1\}$ (with probability one since this is the only possible transition at this step), when the first vintage is created. The process ends at t_1 when there is a single vintage, i.e. $\alpha_1(t_1) = 0$, and $\alpha_2(t_1) = \{n-1\}$. A complete realization of Tajima's coalescent 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 $g^T = \{\boldsymbol{\alpha}_i\}_{i=1}^n$, where $\boldsymbol{\alpha}_i = \{\alpha_{i,1}, \alpha_{i,2}\}$ (Figure 2(c)). The jump chain has the following transition probabilities:

$$P(\boldsymbol{\alpha}_{i-1} \mid \boldsymbol{\alpha}_i) = \begin{cases} \frac{\binom{\alpha_{i,1}}{\alpha_{i,1}-\alpha_{i-1,1}}}{\binom{\alpha_{i,1}+|\alpha_{i,2}|}{2}} & \text{if} \quad \boldsymbol{\alpha}_{i-1} \prec \boldsymbol{\alpha}_i \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

Given (4), 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 partitions joining two singletons (or cherries). We will use $\mathcal{G}_n^T$ to denote the space of such Tajima's topologies.

While Kingman's coalescent keeps track who is related to whom, Tajima's coalescent describes the evolutionary relationships of a sample of n individuals by keeping track 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 in 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 et al. (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.

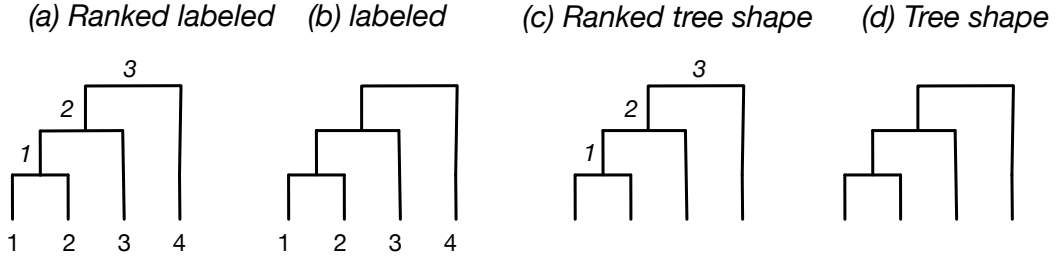

Figure 3: **Tree topologies:** The **(a)** ranked labeled tree topology (Kingman), **(b)** labeled (unranked) tree topology, **(c)** ranked tree shape (Tajima) and **(d)** tree shape

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 $|\mathcal{G}_n^{LT}| = (2n-3)!!$ (Steel 2016); the number of ranked tree shapes $|\mathcal{G}_n^T|$ is the n-th term of the Euler zig-zag sequence (alternating permutations, OEIS: A000111) (Disanto and Wiehe 2013), and the number of tree shapes is the n-th Wedderburn-Etherington number (OEIS: 01190) (Steel 2016).

For $n > 3$, it holds that $|\mathcal{G}_n^{TS}| < |\mathcal{G}_n^{LT}|$ and $|\mathcal{G}_n^T| < |\mathcal{G}_n^K|$. For example, for $n = 5$, there are 180 ranked labeled trees and 5 unlabeled ranked trees. Similarly, 105 labeled trees and 3 tree shapes. This cardinality difference has motivated the study of lower resolution coalescent processes (Sainudiin et al. 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 non-recombining contiguous segment of DNA (or locus) is the result of a mutation process superimposed on the timed genealogy $\mathbf{g}$ (Figure 1). 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). 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. In addition, if the ancestral type at each polymorphic sites 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 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*. 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 (3 groups represented as leaf nodes in Figure 4(a)).

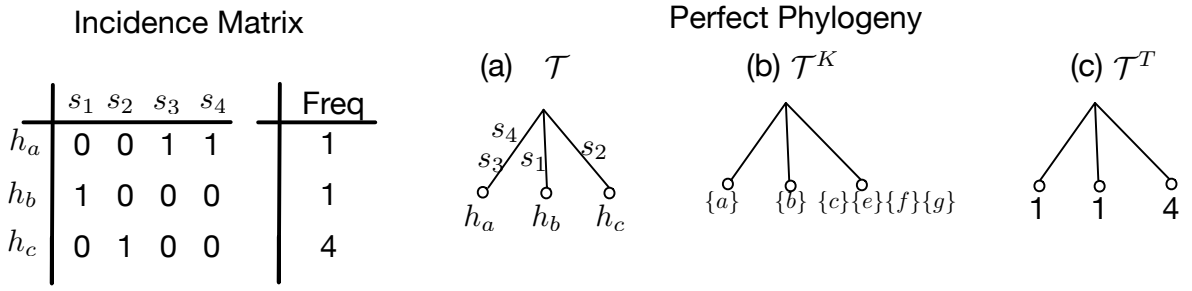

Figure 4: **Perfect phylogeny representation**. Data is summarized as an incidence matrix and a vector of frequencies. **(a)** Original perfect phylogeny $\mathcal{T}$ in bijection with the incidence matrix; each of the 4 polymorphic sites label exactly one edge. When an edge has multiple labels, the order of the labels is irrelevant. Each of the 3 haplotyes 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.

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

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 which 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 individual labels for each haplotype (Figure 4(b)). 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.

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

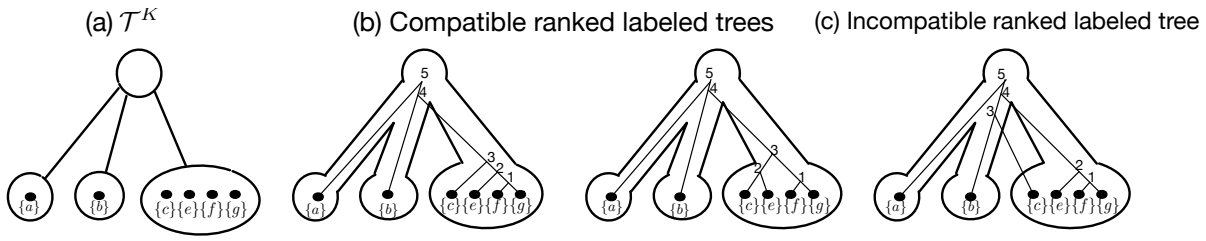

Figure 5: **Compatibility of ranked labeled trees with the perfect phylogeny. (a)** Perfect phylogeny, **(b)** Two examples of ranked labeled trees compatible with the perfect phylogeny, **(c)** incompatible ranked labeled tree.

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, i.e. $|\mathcal{G}_{n,C}|$ is given by

$$\mathrm{E}_q\left[\frac{1}{q(g)}\right] = \sum_{g \in \mathcal{G}_{n,C}} \frac{1}{q(g)} q(g) = |\mathcal{G}_{n,C}|, \tag{5}$$

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

$$\widehat{|\mathcal{G}_{n,C}|} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{q(g_i)}, \tag{6}$$

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

Average (6) is an instance of importance sampling (IS) (Hammersley and Handscomb 1964, 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:

$$q(g) = q(c_n) q(c_{n-1} \mid c_n) \ldots q(c_1 \mid c_2), \tag{7}$$

Approaches with a similar stochastic sequential nature construction have been used for enumeration in other contexts, such as random graphs, networks and contingency tables (Knuth 1976, Chen et al. 2005, Blitzstein and Diaconis 2011, Chen and Chen 2018, Diaconis 2018). It is clear from this literature that the algorithm should satisfy two desiderata: it should not "get stuck", i.e. 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 et al. (2018) show that $N \approx \exp(KL(q,p))$ is necessary

and sufficient for accurate estimation by IS, where KL denotes the Kullback-Leibler divergence. In addition, Chatterjee et al. (2018) warn against the use of sample variance as a criteria 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 ESS $= N/(1 + cv^2)$, and cv^2 is the coefficient of variation given by

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

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 et al. (2018) define $q_N = \text{E}[Q_N]$ where

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

and propose to use a Monte Carlo estimate of q_N below a certain threshold as a criteria 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 both in Q_N and cv^2, so it is possible to compute these two diagnostics.

3.1 Sampling from the perfect phylogeny

To generate a tree topology $g \in \mathcal{G}_{n,C}$ compatible with the observed data $\mathcal{T}$, we proceed sequentially from the tips to the root in both $\mathcal{T}$ and g: one coalescence in g and one node in $\mathcal{T}$ at a time.

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, i.e. nodes with at least two particles; v is an element of V, and $\text{pa}(v)$ 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, i.e. 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 how many individuals have yet to coalesce.

3.1.1 Data constrained Kingman coalescent

To sample a ranked labeled tree $g^K = \{c_i\}_{i=1}^n$ of n individuals compatible with the observed perfect phylogeny $\mathcal{T}^K$, we start at $c_n = \{\{1\}, \ldots, \{n\}\}$. Each leaf node of $\mathcal{T}^K$ contains a subset of c_n, which we denote c_n^v, corresponding to the set of individuals assigned to that node.

The first step is to define the set L: we remove the leaf nodes with a single individual ($|c^v| = 1$) and assign those individuals to their parent nodes. L is the set of nodes with at least two particles. 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 $v \in L$ is $q_i^1(v) = |c_{i+1}^v| / \sum_{j \in L} |c_{i+1}^j|$. The node sampled at iteration i is denoted by v_i. The transition from $c_{i+1}^{v_i}$ to $c_i^{v_i}$ consists in joining two elements

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; *i.e.*

$$q_i^2(c_i^{v_i} \mid c_{i+1}^{v_i}) = \begin{cases} \binom{|c_{i+1}^{v_i}|}{2}^{-1} & \text{if } \; c_i^{v_i} \prec c_{i+1}^{v_i} \\ 0 & \text{otherwise} \end{cases} . \tag{8}$$

Note that at every iteration $c_i = \cup_v c_i^v$. The two probabilities q_i^1 and q_i^2 are all we need to compute the transition probability

$$q(c_i \mid c_{i+1}) = q_i^1(v_i) q_i^2(c_i^{v_i} \mid c_{i+1}^{v_i}),$$

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 (7). We outline our sampling algorithm with the following example and provide the pseudocode in Algorithm 1.

Example. 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 $c_6^{j_0} = \emptyset$. 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 probabilities $(4/6, 3/5, 1/2, 1, 1)$. The coalescent events probabilities q_2 are $(1/6, 1/3, 1, 1, 1)$. Thus, $q(g^K) = 1/90$.

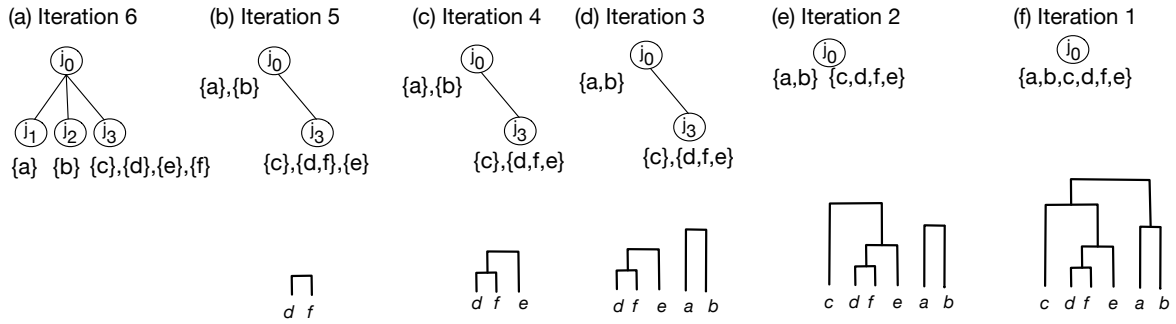

Figure 6: **Example of sequential sampling of a Kingman tree topology with constraints.** We start with a perfect phylogeny (a), at each iteration (b)-(f) we select a node and coalesce a pair from the selected node. The algorithm terminates when a single tree topology of size n is generated.

3.1.2 Data constrained Tajima coalescent

To sample a ranked tree shape $g^T = \{\boldsymbol{\alpha}_i\}_{i=1}^n$ of n individuals compatible with the observed perfect phylogeny $\mathcal{T}^T$ (Figure 4 (c)), we start at $\boldsymbol{\alpha}_n = (n, \emptyset)$ and each leaf node in the perfect phylogeny $\mathcal{T}^T$ is assigned a vector $\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{v}} = (\alpha_{n,1}^v, \alpha_{n,2}^v)$. Recall that $\alpha_{n,1}^v$ denotes

Algorithm 1 Sequential sampling on a constrained Kingman tree topology

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

1. If a leaf node v is such that $|c_n^v| = 1$, then we let $c_n^{\text{pa}(v)} = c_n^{\text{pa}(v)} \cup c_n^v$ and $c_n^v = \emptyset$.
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_i^1.
   (b) Choose particles in v_i to coalesce with probability q_i^2.
   (c) Update $c_{i-1}^{v_i}$ and define $c_{i-1}^v = c_i^v$ for all the other nodes.
   (d) If $|c_{i-1}^{v_i}| = 1$, we let $c_{i-1}^{\text{pa}(v_i)} = c_{i-1}^{\text{pa}(v_i)} \cup c_{i-1}^{v_i}$ and $c_{i-1}^{v_i} = \emptyset$.
   (e) Update $q = q \times q_i^1 \times q_i^2$
   (f) Update L as the list of nodes such that $|c_{i-1}^v| > 1$
5. **end for**

the number of singletons, and $\alpha_{n,2}^v$ denotes the set of vintages associated to node v. Initially, each leaf node in the perfect phylogeny contains the number of singleton particles $\sum_{v\in V} \alpha_{n,1}^v = n$, and no vintages, i.e. $\alpha_{n,2}^v = \emptyset$ for all $v \in V$. At any given iteration i, the number of particles associated to a node v is $\alpha_{i,1}^v + |\alpha_{i,2}^v|$.

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 $q_i^1(v) = (\alpha_{i,1}^v + |\alpha_{i,2}^v|)/\sum_{j\in L}(\alpha_{i,1}^j + |\alpha_{i,2}^j|)$; then we sample a pair of particles in the selected node to coalesce. Our proposal q_i^2 is:

$$q_i^2(\boldsymbol{\alpha_i^{v_i}} \mid \boldsymbol{\alpha_{i+1}^{v_i}}) = \begin{cases} \binom{\alpha_{i+1,1}^{v_i}}{\alpha_{i+1,1}^{v_i} - \alpha_{i,1}^{v_i}} \binom{\alpha_{i+1,1}^{v_i} + |\alpha_{i+1,2}^{v_i}|}{2}^{-1} & \text{if } \boldsymbol{\alpha_i^{v_i}} \prec \boldsymbol{\alpha_{i+1}^{v_i}} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Analogously to the Kingman sampler, each iteration ends by updating $\boldsymbol{\alpha_i^{v_i}}$ and L. The pseudocode is presented in Algorithm 2. Note that, as opposed to the Kingman sampler, q_i^1 and q_i^2 in Tajima sampling, do not fully determine $q(\boldsymbol{\alpha_i}|\boldsymbol{\alpha_{i+1}})$, where $\boldsymbol{\alpha_i} = (\sum_{v\in V} \alpha_{i,1}^v, \bigcup_{v\in V} \alpha_{i,2}^v)$. A transition from $\boldsymbol{\alpha_{i+1}}$ to $\boldsymbol{\alpha_i}$ 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 $\{\boldsymbol{\alpha}_i\}_{i=1}^n = g^T$ 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. 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 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 4 possible sampling

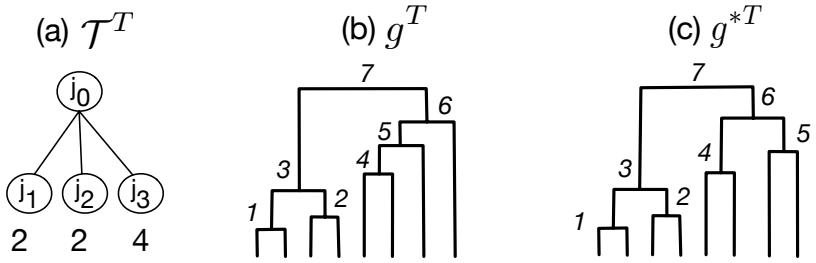

Figure 7: **Example of two ranked tree shapes compatible with a given perfect phylogeny.** (a) perfect phylogeny (b)-(c) two possible ranked tree shapes compatible with $\mathcal{T}^T$.

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 $|\mathcal{G}^T_{n,c}|$.

Definition 3. Let $\mathcal{Y}^T_{n,C}$ 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. Let $c^T(g^T)$ the number of possible pairs $(g^T, \boldsymbol{v'}) \in \mathcal{Y}^T_{n,C}$ equivalent to $(g^T, \boldsymbol{v})$.

It is still possible to use sequential importance sampling despite the fact that our proposal q has support $\mathcal{Y}^T_{n,C}$ instead of $\mathcal{G}^T_{n,C}$. We discuss two alternative ways. The first one is to generate a sample $(g^T, \boldsymbol{v}) \in \mathcal{Y}^T_{n,C}$ with sampling probability $q(g^T, \boldsymbol{v})$ computed as the product of all q^1 and q^2 transition probabilities (Algorithm 2). We then call a depth-first search algorithm that backtracks all possible sequence of nodes $\boldsymbol{v'}$ that would give rise to the same g^T and compute:

$$q(g^T) = \sum_{\boldsymbol{v'}:(g^T,\boldsymbol{v'})\in\mathcal{Y}^T_{n,C}} q(g^T, \boldsymbol{v'}). \tag{10}$$

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

$$\begin{aligned} E_{\mathcal{Y}^T_{n,C}}\left[\frac{1}{q(g^T)}\right] &= \sum_{(g^T,\boldsymbol{v})\in\mathcal{Y}^T_{n,C}} \frac{q(g^T,\boldsymbol{v})}{q(g^T)} = \sum_{g^T\in\mathcal{G}^T_{n,C}} \frac{1}{q(g^T)} \sum_{\boldsymbol{v}:(g^T,\boldsymbol{v})\in\mathcal{Y}^T_{n,C}} q(g^T,\boldsymbol{v}) \\ &= \sum_{g^T\in\mathcal{G}^T_{n,C}} \frac{q(g^T)}{q(g^T)} = |\mathcal{G}^T_{n,C}| \end{aligned} \tag{11}$$

Our implementation is based on (11), however we discuss a second alternative for completeness. This second alternative is inspired by a similar situation discussed in Blitzstein and Diaconis (2011) in the context of sampling graphs with a given degree sequence. The cardinality is estimated by the Monte Carlo approximation to the following

$$\begin{aligned} E_{\mathcal{Y}^T_{n,C}}\left[\frac{1}{c^T(g^T)q(g^T,\boldsymbol{v})}\right] &= \sum_{(g^T,\boldsymbol{v})\in\mathcal{Y}^T_{n,C}} \frac{q(g^T,\boldsymbol{v})}{c^T(g^T)q(g^T,\boldsymbol{v})} \\ &= \sum_{g^T\in\mathcal{G}^T_{n,C}} \frac{1}{c^T(g^T)} \sum_{\boldsymbol{v}:(g^T,\boldsymbol{v})\in\mathcal{Y}^T_{n,C}} 1 = |\mathcal{G}^T_{n,C}|, \end{aligned} \tag{12}$$

where $c^T(g^T)$ is the size of the equivalence class $c^T(g^T) = \#\{\boldsymbol{v'} : (g^T, \boldsymbol{v'}) \in \mathcal{Y}^T_{n,C}\}$.

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 $\mathcal{T}^T$. More specifically, let $g^T_{v_i}$ be the subtree in g^T created at iteration i. We say that two subtrees $g^T_{v_i}$ and $g^T_{v_j}$ are equivalent if they have the same tree shape and $d(v_i, v_j) \leq 2$, where $d(v_i, v_j)$ is the number of edges separating v_i and v_j in $\mathcal{T}^T$. We only consider equivalence classes of subtrees $g^T_{v_i}$ if v_i is either an internal node in $\mathcal{T}^T$ or a leaf node in $\mathcal{T}^T$ removed from L at iteration i. Let K be number of such equivalence classes and let $E_1, \ldots, E_K$ denote the equivalence classes of subtrees of g^T, then

$$c^T(g^T) = \prod_{k=1}^{K} |E_k|! \tag{13}$$

One can see that formula (13) is computationally expensive and efficient implementation of (13) is an open question.

Algorithm 2 Sampling on the constrained Tajima Space

Inputs: $\mathcal{T}^T$, with $\alpha^v_{n,1}$ number of singletons at all leaf nodes, and $\alpha^v_{n,2} = \emptyset$ for all $v \in V$.
Outputs: g^T, $q(g^T)$

1. If a leaf node v is such that $\alpha^v_{n,1} = 1$, then let $\alpha^{pa(v)}_{n,1} = \alpha^{pa(v)}_{n,1} + 1$, $\alpha^v_{n,1} = 0$.
2. Define L as the list of nodes with $\alpha^v_{n,1} > 1$.
3. **for** $i = n - 1$ to 1 **do**
   (a) Sample node v_i with probability q^1_i.
   (b) Choose particles to coalesce with probability q^2_i
   (c) Update $\boldsymbol{\alpha^{v_i}_{i-1}}$ and define $\boldsymbol{\alpha^{v_i}_{i-1}} = \boldsymbol{\alpha^{v_i}_{i}}$ for all other nodes
   (d) If $\alpha^{v_i}_{i-1,1} + |\alpha^{v_i}_{i-1,2}| = 1$, then let $\alpha^{pa(v_i)}_{i-1,1} += \alpha^{v_i}_{i-1,1}$, $\alpha^{v_i}_{i,1} = 0$, and $\alpha^{pa(v_i)}_{i-1,2} \cup = \alpha^{v_i}_{i-1,2}$, $\alpha^{v_i}_{i,2} = \emptyset$
   (e) Update $q = q \times q^1_i \times q^2_i$
4. **end for**
5. Compute $q(g^T)$ as in (10).

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 $\mathcal{G}^K_{n,C}$ and $\mathcal{G}^T_{n,C}$ into equivalence classes that ignore rankings. We show two simple formulas to compute the size of these classes. Opposed to (13), these formulas are easy to implement and allow to build a SIS procedure to estimate $|\mathcal{G}^{LT}_{n,C}|$ and $|\mathcal{G}^{TS}_{n,C}|$ 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 $g^K \in \mathcal{G}^K_{n,C}$, let $LT(g^K)$ denote the corresponding unranked labeled tree $g^{LT} \in \mathcal{G}^{LT}_{n,C}$. We call g^K and g'^K equivalent if $LT(g^K) = LT(g'^K)$ and we denote the size of the equivalence class by $c^{LT}(g^K)$.

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

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

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

Proof. Note that $|g^K_{i,j}|-1$ is the number of coalescent events in subtree $g^K_{i,j}$. For each fixed i, we are computing the number of possible permutations of $(|g^K_{i,1}|+|g^K_{i,2}|-2)$ coalescent events of elements of two groups with $|g^K_{i,1}|-1$ and $|g^K_{i,2}|-1$ elements respectively. The two groups are guaranteed by the fact that g^K is a binary tree. The product accounts for all possible orderings. □

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

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

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

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

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 dived by two since the same rankings in the two subtrees are indistinguishable. □

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

4 Applications and simulations

4.1 Simulation studies

In this section we discuss two simulation studies. First, we assess the convergence of our algorithms and discuss the diagnostics employed. We then analyze and quantify the differences in cardinalities between the four tree topologies for a given simulated dataset. The four tree topologies analyzed are $\mathcal{G}^K_{n,C}$: Kingman ranked labeled trees, $\mathcal{G}^T_{n,C}$: Tajima ranked tree shapes, $\mathcal{G}^{TS}_{n,C}$: tree shapes and $\mathcal{G}^{LT}_{n,C}$: unranked labeled trees. All of which are compatible with the simulated dataset.

To simulate a single dataset, we first simulate a Kingman genealogy of n individuals as described in Section 2 and implemented in `R ape:rcoal()` (Paradis et al. 2004). We assume a constant effective population size and thus the k-th coalescent time is exponential distributed with rate $\binom{k}{2}$. Given a timed genealogy with tree length $L = \sum_{k=2}^{n} kt_k$, a number m of mutations is sampled as a Poisson random variable with rate μL. These m mutations are then placed uniformly at random along the branches of the timed genealogy and labeled $1,\ldots,m$. The resulting incidence matrix is initially a matrix of size $n \times m$ with (i,j) entry equal to 1 if the branch path from leaf i to the root has labeled mutation j. This part of the simulation algorithm corresponds the infinite-sites mutation model. The final incidence matrix is then summarized by unique rows (haplotypes).

Lacking a competing method for accuracy check, we resort to two type of checks. First, we use the SIS diagnostics described in Section 3: coefficient of variations cv^2, effective sample size ESS, and standard error se. Second, for $n \leq 10$, we estimate the cardinality by the number of observed distinct topologies from 3×10^5 $i.i.d$ trees. This second "brute force" method becomes computationally unfeasible already at small n. We refer to it as "real count".

4.1.1 Numerical convergence

We simulate incidence matrices under four scenarios: with sample sizes $n \in (10, 20)$, and two mutation regimes $\mu \in (5, 20)$. SIS estimates and diagnostics are computed at $N \in (100, 500, 1000, 3000, 5000, 10000, 15000)$ from 20 repetitions on each of the four incidence matrices. Figure 8 shows the results for the four combinations of n and μ (columns) and the four topological spaces considered (rows). Grey lines represent estimates as N increases, the black line denotes the mean estimate.

A visual inspection of Figure 8 suggests a few qualitative observations. For $n = 10$, convergence is generally achieved for fairly small SIS sample size ($N \approx 3000$), although variance across runs decreases solely at $N = 10000$. As expected, larger sample sizes (n) increase estimation uncertainty; this is depicted by scattered and non overlapping grey lines; in general, the coefficient of variation increases one order of magnitude by increasing the sample size n from 10 to 20. Similarly, estimates of unranked topologies: tree shapes and labeled trees show higher coefficients of variation than the ranked counterparts. This is not surprising given that the sampling distributions for $\mathcal{G}^{LT}_{n,C}$ and $\mathcal{G}^{TS}_{n,C}$ are "corrected" from the their corresponding ranked tree spaces. A comparison between mean squared coefficients of variation (computed for $N > 5000$) show that Kingman's algorithm performs better than Tajima's algorithm. Tajima's algorithm mean cv^2 are 0.87 and ≈ 7.87 for $n = 10$ and $n = 20$ respectively, while Kingman's algorithm mean cv^2 are 18% and 40% lower.

4.1.2 Multi-resolution simulation study

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

n	μ	$\lvert J_{leaf}\rvert$	$\lvert J\rvert$	Tajima trees	Kingman trees	Tree shapes	Labeled trees
5	5	2	4	2.994 +/- 0.01	9.01 +/- 0.01223	3.003 +/- 0.004076	0.9979 +/- 0.003333
	10	2	5	3.011 +/- 0.01	9.018 +/- 0.01221	3.006 +/- 0.00407	1.004 +/- 0.003333
	20	2	5	3.005 +/- 0.01	9.018 +/- 0.01221	3.006 +/- 0.00407	1.002 +/- 0.003333
	50	3	7	3.009 +/- 0.01	3 +/- 0.01	0.9999 +/- 0.003333	1.003 +/- 0.003333
10	5	5	8	1410 +/- 14.66	10520 +/- 101	52.73 +/- 1.057	8.478 +/- 0.2009
	10	7	14	361.5 +/- 3.407	724.9 +/- 6.587	3.329 +/- 0.03942	2.306 +/- 0.03901
	20	7	14	360.9 +/- 3.378	725.2 +/- 6.589	3.329 +/- 0.03941	2.338 +/- 0.03874
	50	7	16	281.5 +/- 2.368	282.8 +/- 2.358	1.01 +/- 0.008423	1.006 +/- 0.008456
15	5	7	14	9532000 +/- 186600	2.65e+08 +/- 4909000	1961 +/- 52.38	113.8 +/- 2.901
	10	7	16	504100 +/- 7710	9458000 +/- 124300	48.18 +/- 0.617	4.175 +/- 0.05833
	20	7	16	526700 +/- 8191	9467000 +/- 124400	48.19 +/- 0.6165	4.293 +/- 0.06187
	50	8	20	256100 +/- 3514	1578000 +/- 19390	9.962 +/- 0.1224	3.234 +/- 0.04437
20	5	8	17	2.486e+09 +/- 86600000	9.007e+12 +/- 2.574e+11	241900 +/- 17940	101.4 +/- 8.59
	10	9	20	2.063e+09 +/- 56090000	6.195e+12 +/- 1.503e+11	140900 +/- 9516	72.43 +/- 4.551
	20	11	23	1.306e+09 +/- 50380000	5.668e+10 +/- 2.084e+09	540.1 +/- 59.65	24.18 +/- 2.135
	50	11	26	462900000 +/- 14690000	1.024e+10 +/- 337900000	39.4 +/- 2.042	4.056 +/- 0.1785

We simulate 50 incidence matrices for the 20 possible pairings of n in $(5, 10, 15, 20)$ and μ in $(5, 10, 20, 50, 75)$. 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 $n \in (5, 10)$, $N = 10000$ for $n = 15$, and $N = 15000$ for $n = 20$. In the first row of Figure 9, we show the log ratio of the estimated cardinalities of Kingman topologies and Tajima topologies. Similarly, in the second row of Figure 9, we show the log ratio of the estimated cardinalities of labeled trees and tree shapes. Table 1 summarizes the results for a single iteration picked at random among the 50 replicates: an average

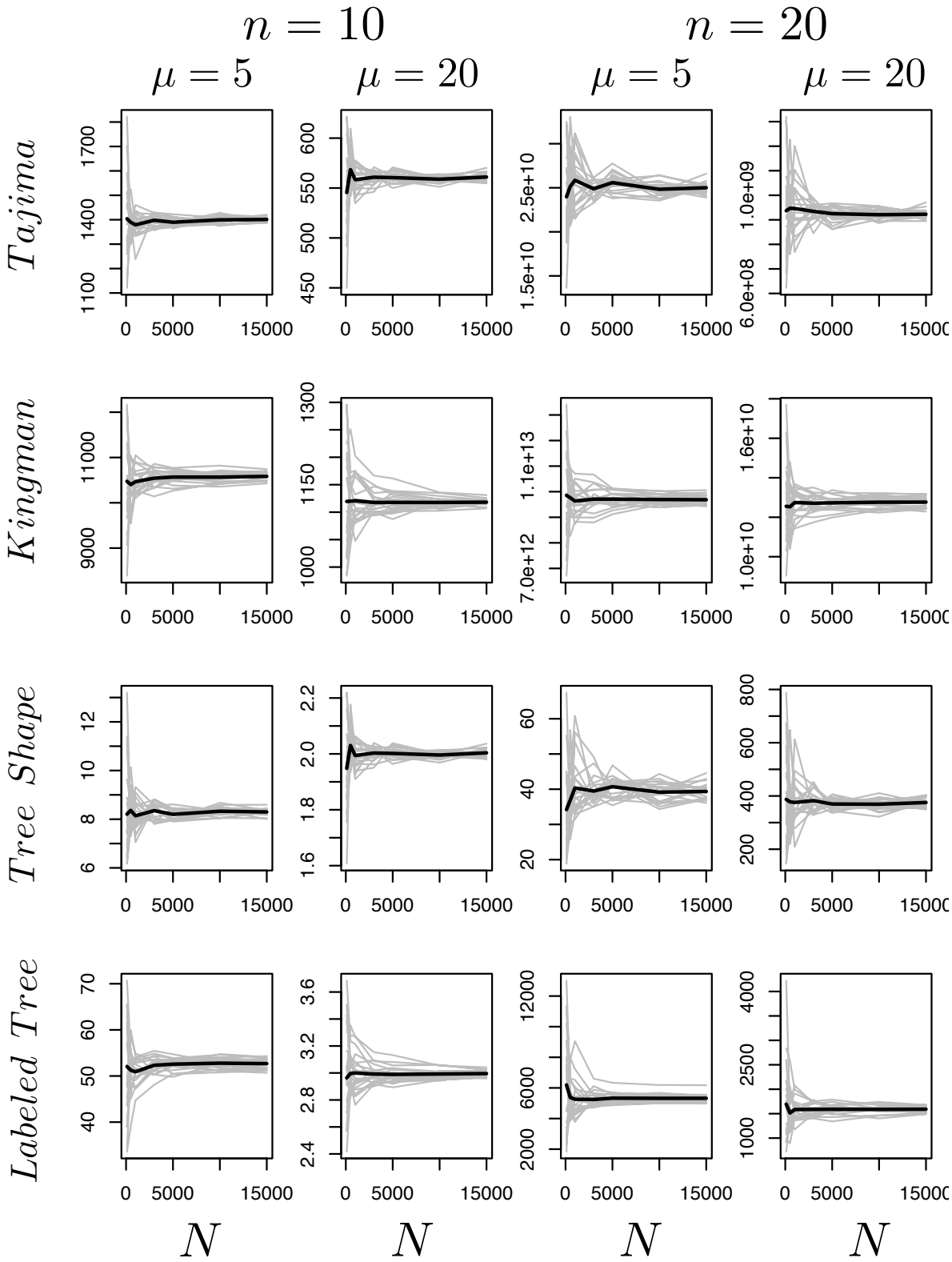

Figure 8: **Sequential importance sampling convergence.** Rows show the four topologies: Ranked tree shapes, ranked labeled trees, unranked tree shapes and labeled tress respectively, the first two columns show results on simulations based on $n = 10$ samples and the last two columns on $n = 20$ samples. The first and third column show results for $\mu = 5$ and second and fourth show results for $\mu = 20$. Grey lines show each of the 20 independent estimates from the 20 repetitions of the SIS algorithm computed at $N \in (100, 500, 1000, 3000, 5000, 10000, 15000)$ iterations. Black lines show the mean estimate of the 20 repetitions.

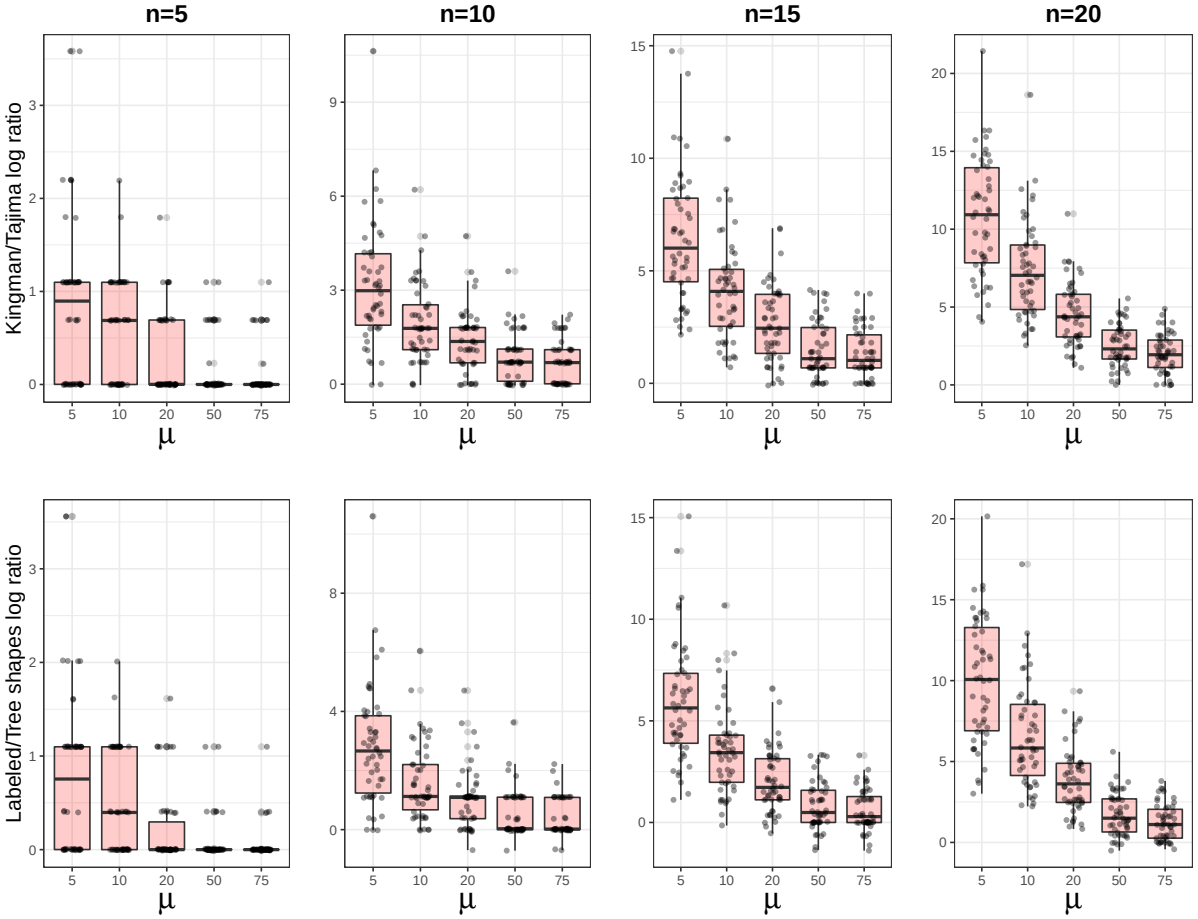

Figure 9: **Log ratio of estimated counts for varying n and μ.** Rows correspond to the log ratio of cardinalities between Kingman and Tajima topologies (first row) and the log ratio of cardinalities between labeled trees and tree shapes (second row). Columns represent different sample sizes n and boxplots 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 $n = 5, 10$), $N = 10000$ (for $n = 15$), and $N = 15000$ (for $n = 20$). Dots are spread over the box width for ease of visualization.

over the 50 iterations is not insightful given the high variability of the incidence matrices sampled (which can be observed in Figure 9).

The reduction in cardinality of Tajima's versus Kingman's is clearly depicted in Figure 9, as it is for the reduction in cardinality of tree shapes versus labeled trees. However, as the rate of mutation μ increases, these two ratios become exponentially smaller. If we compare these ratios as the sample size n increases, we also observe an exponential growth trend (in natural scale), highlighting the difference in cardinalities between unlabeled and labeled trees.

Recall that, in expectation, a higher mutation rate corresponds to a higher number of mutations, which should correspond to a more constrained space. The reduction in the tree spaces imposed by mutations is accentuated for Kingman's trees under every 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.

Lastly, the benefits of employing coarser resolutions are striking when the sample sizes increases. For $n = 20$ and $\mu = 5$, the Tajima space is on average (across the 50 datasets) 48000 times smaller than the Kingman space: 60 times smaller in the simulated worst case scenario, and two billion times smaller in the best case.

4.2 Human mtDNA data

The left of Figure 10 shows the Tajima perfect phylogeny reconstructed from $n = 35$ 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. 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).

We estimated the cardinalities of the four constrained topologies by running ten iterations of the algorithms for $N \in (5000, 10000, 15000, 20000, 25000)$ iterations. Computing SIS estimates for increasing N allows to check the convergence. We report mean estimate across the 10 for N larger or equal 15000. Note that in this way, we are implicitly increasing N ten fold.

The cardinality of the unconstrained Kingman space (for $n = 35$) is 1.78×10^{68} trees, while our estimated cardinality of compatible Kingman's trees is $3.65 \times 10^{29} \pm 2.65 \times 10^{28}$. The unconstrained cardinality of Tajima space is 8.07×10^{31}, while our estimated cardinality of compatible Tajima space is $\mathbf{1.81 \times 10^{25}} \pm 1.96 \times 10^{24}$. We show that the space of compatible Tajima's trees if 4 orders of magnitude smaller than the space of compatible Kingman's tree. While Tajima's constrained space has the smallest cardinality, we note that in this setting with high mutation rates and, as pointed out in the introduction, inference would be computationally expensive.

Moving now to the unranked tree spaces, the cardinality of unconstrained labeled tree space is 4.89×10^{47}, while the estimated cardinality of compatible labeled tree space is $9 \times 10^{16} \pm 4.91 \times 10^{16}$. Finally, the constrained cardinality of tree shapes is $2.06 \times 10^{13} \pm 1.05 \times 10^{13}$ (unconstrained cardinality unavailable). We note the very high standard errors of the unranked estimates may be a product of low algorithmic efficiency for unranked trees.

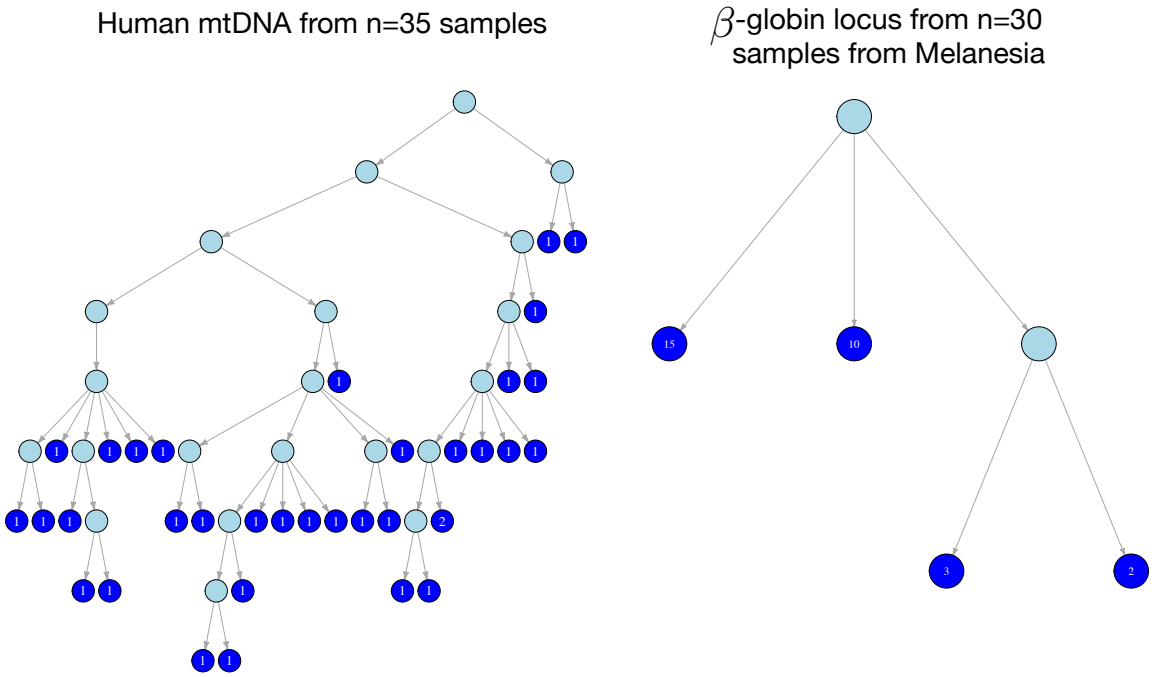

Figure 10: **Tajima perfect phylogenies of Yoruban mitochondrial data (left) and Melanesian $\beta-$globin locus data (right).** Left panel: $\mathcal{G}^T$ of $n = 35$ sequences of mtDNA sampled at random from 107 Yoruban individuals available from the 1000 Genomes Project phase 3 (1000 Genomes Project Consortium 2015). Right panel: $\mathcal{G}^T$ of $n = 30$ sequences of DNA from the β-globin locus sampled at random from 57 Melanesian individuals available in Fullerton et al. (1994). Dark blue nodes represent the leaf nodes. The number within a node is the number of individuals assigned to that node.

4.3 Melanesian β-globin locus

The right side of Figure 10 shows the reconstructed Tajima perfect phylogeny from part of the β-globin locus. This dataset consist of $n = 30$ sequences sampled uniformly at random from the 57 sequences from a Melanesian human population analyzed in Fullerton et al. (1994). This dataset was already part of a larger dataset described in Harding et al. (1997)and consist of 13 segregating sites and 4 haplotypes (distinct sequences among the 30 sequences).

Similar to our previous analysis on mtDNA data, we run ten iterations of the algorithm. However, in this case we only estimate cardinalities of the two ranked type of trees: Kingman and Tajima. For unranked trees, it is not possible to assert algorithmic convergence for already high N; this is not a complete surprise since we have already pointed out the poor performance of estimates for unranked tree spaces. The cardinality of the unconstrained Kingman space (for $n = 30$) is 4.37×10^{54} while the constrained Kingman space has estimated cardinality of $1.07 \times 10^{40} \pm 6.66 \times 10^{37}$. The cardinality of the unconstrained Tajima space is 2.31×10^{25}, while the constrained space has cardinality $2.98 \times 10^{23} \pm 5.01 \times 10^{22}$.

5 Discussion

In this article we propose a set of algorithms to sample coalescent tree topologies when the infinite sites mutation model is assumed. Our algorithms sequentially sample tree topologies compatible with the observed data in order to estimate their cardinality using importance sampling. We analyze the cardinality of different types of tree topologies: ranked labeled trees (Kingman), ranked tree shapes (Tajima), unranked labeled trees and tree shapes, which corresponds 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. In our application either Kingman or Tajima coalescent process.

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 being clear that there is a reduction in the cardinality of the state-space using coarser resolutions, e.g. Tajima n-coalescent, there was no way to quantify how much smaller the state space is conditionally on the data. 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 a different resolution before any more work is carried out.

From our empirical analysis, it emerges that the benefits of using a coarser resolution depends largely on the data considered. The advantages are striking as the sample size increases, especially in the context of few segregating sites (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: 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 mutation rate is low, and consequently the number of mutations is also low, the benefits of lower resolutions remain clear.

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.

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., Sanger, F. et al. (1981), 'Sequence and organization of the human mitochondrial genome', Nature **290**(5806), 457.

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

Behar, D. M., Van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N. M., Kivisild, T., Torroni, A. and Villems, R. (2012), 'A Copernican reassessment of the human mitochondrial DNA tree from its root', The American Journal of Human Genetics **90**(4), 675–684.

Blitzstein, J. and Diaconis, P. (2011), 'A sequential importance sampling algorithm for generating random graphs with prescribed degrees', Internet mathematics **6**(4), 489–522.

Chatterjee, S., Diaconis, P. et al. (2018), 'The sample size required in importance sampling', The Annals of Applied Probability **28**(2), 1099–1135.

Chen, Y. and Chen, Y. (2018), 'An efficient sampling algorithm for network motif detection', Journal of Computational and Graphical Statistics **0**(0), 1–13.

Chen, Y., Diaconis, P., Holmes, S. P. and Liu, J. S. (2005), 'Sequential Monte Carlo methods for statistical analysis of tables', Journal of the American Statistical Association **100**(469), 109–120.

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', Mathematical biosciences **242**(2), 195–200.

Drummond, A., Suchard, M., Xie, D. and Rambaut, A. (2012), 'Bayesian phylogenetics with BEAUti and the BEAST 1.7', Molecular Biology and Evolution **29**(8), 1969–1973.

Fullerton, S., Harding, R., Boyce, A. and Clegg, J. (1994), 'Molecular and population genetic analysis of allelic sequence diversity at the human beta-globin locus.', Proceedings of the National Academy of Sciences **91**(5), 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**(1), 235–245.

Gattepaille, L., Günther, T. and Jakobsson, M. (2016), 'Inferring past effective population size from distributions of coalescent times', Genetics **204**(3).

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

Hammersley, J. and Handscomb, D. (1964), Monte Carlo Methods, Methuen and Co, Ltd, London.

Harding, R., Fullerton, S., Griffiths, R. and Clegg, J. (1997), 'A gene tree for β-globin sequences from Melanesia', Journal of Molecular Evolution **44**(1), 133–138.

Kimura, M. (1969), 'The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations', Genetics **61**(4), 893.

Kingman, J. F. (1982), 'On the genealogy of large populations', Journal of applied probability **19**(A), 27–43.

Knuth, D. E. (1976), 'Mathematics and computer science: coping with finiteness.', Science (New York, NY) **194**(4271), 1235–1242.

Liu, D., Shi, W., Shi, Y., Wang, D., Xiao, H., Li, W., Bi, Y., Wu, Y., Li, X., Yan, J. et al. (2013), 'Origin and diversity of novel avian influenza a h7n9 viruses causing human infection: phylogenetic, structural, and coalescent analyses', The Lancet **381**(9881), 1926–1932.

Nordborg, M. (1998), 'On the probability of Neanderthal ancestry.', American journal of human genetics **63**(4), 1237.

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**(1), 8–18.

Palacios, J. A., Véber, A., Cappello, L., Wang, Z., Wakeley, J. and Ramachandran, S. (2019+), 'Bayesian estimation of population size changes by sampling Tajimas trees', in preparation .

Paradis, E., Claude, J. and Strimmer, K. (2004), 'Ape: analyses of phylogenetics and evolution in r language', Bioinformatics **20**(2), 289–290.

Rosenberg, N. A. and Nordborg, M. (2002), 'Genealogical trees, coalescent theory and the analysis of genetic polymorphisms', Nature Reviews Genetics **3**(5), 380.

Sainudiin, R., Stadler, T. and Véber, A. (2015), 'Finding the best resolution for the Kingman–Tajima coalescent: theory and applications', Journal of Mathematical Biology **70**(6), 1207–1247.

Sainudiin, R. and Véber, A. (2018), 'Full likelihood inference from the site frequency spectrum based on the optimal tree resolution', Theoretical Population Biology **124**, 1–40.

Steel, M. (2016), Phylogeny: discrete and random processes in evolution, SIAM.

Tajima, F. (1983), 'Evolutionary relationship of dna sequences in finite populations', Genetics **105**(2), 437–460.

Terhorst, J., Kamm, J. A. and Song, Y. S. (2017), 'Robust and scalable inference of population history from hundreds of unphased whole genomes', Nature Genetics **49**(2), 303–309.

Wang, L., Bouchard-Côté, A. and Doucet, A. (2015), 'Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method', Journal of the American Statistical Association **110**(512), 1362–1374.