Finding the best resolution for the Kingman-Tajima coalescent: theory and applications

R. Sainudiin · T. Stadler · A. Véber

Abstract Many summary statistics currently used in population genetics and in phylogenetics depend only on a rather coarse resolution of the underlying tree (the number of extant lineages, for example). Hence, for computational purposes, working directly on these resolutions appears to be much more efficient. However, this approach seems to have been overlooked in the past.

In this paper, we describe six different resolutions of the Kingman-Tajima coalescent together with the corresponding Markov chains, which are essential for inference methods. Two of the resolutions are the well-known n-coalescent and the lineage death process due to Kingman. Two other resolutions were mentioned by Kingman and Tajima, but never explicitly formalized. Another two resolutions are novel, and complete the picture of a multi-resolution coalescent. For all of them, we provide the forward and backward transition probabilities, the probability of visiting a given state as well as the probability of a given realization of the full Markov chain. We also provide a description of the state-space that highlights the computational gain obtained by working with lower-resolution objects. Finally, we give several examples of sum-

Raazesh Sainudiin
Biomathematics Research Centre and Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch 8041, New Zealand
E-mail: r.sainudiin@math.canterbury.ac.nz

Tanja Stadler
ETH Zürich, Institut f. Integrative Biologie
CHN H 72, Universitätstrasse 16
8092 Zürich, Switzerland
E-mail: tanja.stadler@env.ethz.ch

Amandine Véber
Centre de Mathématiques Appliquées
École Polytechnique
Route de Saclay
91128 Palaiseau Cedex, France
E-mail: amandine.veber@cmap.polytechnique.fr

mary statistics that depend on a coarser resolution of Kingman's coalescent, on which simulations are usually based.

Keywords n-coalescent resolutions, tree shape statistics, computationally efficient and statistically sufficient inference.

Mathematics Subject Classification (2000) 92D15 · 92D20 · 60J10

1 Introduction

Kingman's n-coalescent (Kingman, 1982a,b) is a process of central importance in mathematical population genetics. It describes the genealogical relations between a sample of n individuals in an infinite population evolving according to the neutral Wright-Fisher model (Fisher, 1930; Wright, 1931). In addition, its robustness to small perturbations of the reproduction mechanism makes it a rather "universal" model of genealogies, in that it appears in fact in many other situations, including, populations with selfing, or evolving in a fluctuating environment. See for instance Etheridge (2011, Sects. 2.2 and 2.3).

The n-coalescent is a continuous-time Markov chain taking its values in the set $\mathbb{C}_n$ of partitions of the label set $\mathfrak{L} = \{1, 2, \ldots, n\}$: at time t, each block contains the labels of individuals sampled at time 0 which have a common ancestor t units of time in the past. The merger of several blocks at some time t' hence means that at time t' in the past, the ancestors corresponding to these blocks have a common parent, and so all the individuals labeled by an element of these blocks find their most recent common ancestor. The n-coalescent thus starts at $\{\{1\}, \ldots, \{n\}\}$. We assume that only a single pair of blocks can merge at any given time, and that each pairwise merger happens at rate 1. The process stops when it has reached its absorbing state $\{\{1, \ldots, n\}\}$ (i.e., the most recent common ancestor of the whole sample has been found). If one considers just the discrete skeleton or the embedded jump chain of this Markov chain, then at each time step one picks two blocks of the partition at random and merges them together, until there is just a single block after $n - 1$ time steps. Unless explicitly specified, from now on we work with this discrete skeleton only (but see Sects. 1.2 and 4 for the introduction of the time component in inference methods).

In this paper, we consider six variants or genealogical resolutions of the discrete n-coalescent process. They are briefly introduced below.

- The *vintaged and labeled* n-coalescent $\{B^{\uparrow}(t)\}$ of Sect. 3.2 is the same as the process described above except that, at all times, each block of the partition has an associated number called the *vintage*, which records the time step or coalescent epoch in which the block was created. Its state space $\mathbb{B}_n$ is an augmentation of $\mathbb{C}_n$ with coalescent vintage tags. This is the Kingman-Tajima n-coalescent.
- The *unvintaged and labeled* n-coalescent $\{C^{\uparrow}(t)\}$ of Sect. 3.3 is obtained from $\{B^{\uparrow}(t)\}$ by dropping the vintages. This is the standard Kingman n-coalescent. Every sequence of states in $\mathbb{C}_n$ that is visited by this process is an element of $\mathscr{C}_n$, the set of n-coalescent sequences or c-sequences. A c-sequence induces a *ranked, rooted, binary tree* (Definition 1) with leaves labeled by $\mathfrak{L}$ and depicted in Fig. 2.

- The *vintaged and sized* n-coalescent $\{D^{\uparrow}(t)\}$ of Sect. 3.4 is obtained from $\{B^{\uparrow}(t)\}$ by keeping track only of the vintage and the size of each block of the partition, and dropping the integer labels $1,2,\ldots,n$. Its state space $\mathbb{D}_n$ is the space of all ordered integer partitions.
- The *vintaged and shaped* n-coalescent $\{G^{\uparrow}(t)\}$ of Sect. 3.5 is obtained from $\{D^{\uparrow}(t)\}$ by keeping track only of the vintages of the blocks at each time step, and throwing away the sizes of the blocks. The state space $\mathbb{G}_n$ is contained in the vertices of the hypercube $\{0,1\}^{n-1}$. The sequence of states visited by this process gives Tajima's *evolutionary relationships* (Tajima, 1983, Figures 1-4), which resolve genealogical histories up to *ranked, rooted, binary tree shapes*. This is Tajima's n-coalescent.
- The *unvintaged and sized* n-coalescent $\{F^{\uparrow}(t)\}$ of Sect. 3.6 is obtained from $\{C^{\uparrow}(t)\}$ by just keeping track of how many blocks there are of each size. This process is also known as the *label-killed* n-coalescent (Kingman, 1982b, (5.2)) or *unlabeled* n-coalescent (Sainudiin et al, 2011) or *family-size* process (Kendall, 1975; Tavaré, 1983, p. 136-137) on $\mathbb{F}_n$, the integer partitions of n.
- The *pure death* process $\{H^{\uparrow}(t)\}$ of Sect. 3.1 is obtained from any of the other five processes above by just keeping track of the number of blocks or the number of ancestral sample lineages in $\mathbb{H}_n = \{n, n-1, \ldots, 1\}$.

As described above and depicted in Fig. 1, these different resolutions are partially ordered: for example, the *unvintaged and sized n-coalescent* on $\mathbb{F}_n$ can been seen as a coarser resolution of the *unvintaged and labeled n-coalescent* $\{C^{\uparrow}(t)\}$ on $\mathbb{C}_n$ and of the *vintaged and sized n-coalescent* on $\mathbb{D}_n$, since both processes contain all the information needed to describe the evolution of $\{F^{\uparrow}(t)\}$. On the other hand, the *vintaged and shaped* n-coalescent on $\mathbb{G}_n$ is not a coarser resolution of the process $\{C^{\uparrow}(t)\}$ since the latter does not contain the information on the vintages required in the description of $\{G^{\uparrow}(t)\}$.

Here we focus on specific algebraic representations of these six Markov chains and derive their backward-transition, sequence-specific, state-specific and forward-transition probabilities. These derivations are rather straightforward, but to the best of our knowledge they have never explicitly appeared in the literature for all but Kingman's labeled n-coalescent and the pure death process (Kingman, 1982a,b). We believe that a global description of all these resolutions in one place will be useful for their direct application to inference or simulations.

Before we start our program, let us describe further motivations for this study. The first is historical and the second is statistical as outlined in the next two subsections, respectively.

1.1 Historical Motivation

Kingman and Tajima independently described the genealogical or evolutionary relationship of a sample of size n from a Wright-Fisher population in the early 1980s. The relation between the genealogical objects introduced by Kingman and Tajima has not been characterized before. Moreover, Kingman's description has come to dominate the literature at the expense of overlooking those of biologists such as Tajima. Here

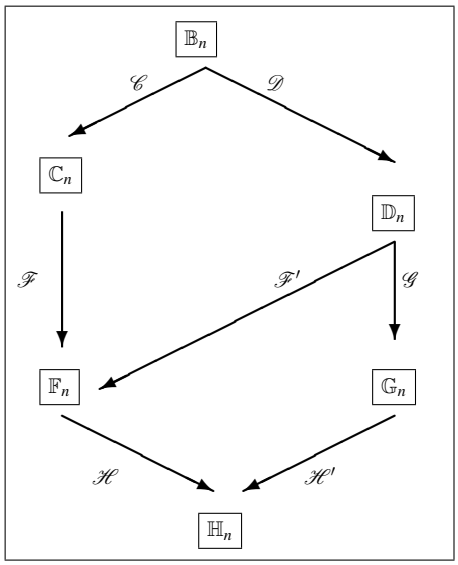


Fig. 1 State spaces $\mathbb{B}_n, \mathbb{C}_n, \mathbb{D}_n, \mathbb{F}_n, \mathbb{G}_n, \mathbb{H}_n$ and the relations between them.

we make the first formalization that connects the two approaches and complete the picture by presenting finer and coarser resolutions that may be of interest in many applications.

In another direction, phylogenetics and population genetics, despite being subfields of mathematical and statistical genetics, are studied by research communities that do not entirely overlap. This is partly driven by methodological preferences between inter-species and intra-species approaches to the study of genetic interrelatedness. This paper attempts to use definitions and notions that are consistent across phylogenetic and population genetic literature in order to facilitate research at the interface of these two historically distinct fields of theoretical evolutionary genetics.

For instance, we show how different resolutions of coalescent sequences are in bijection with different kinds of phylogenetic trees. We also show that classical phylogenetic tree shape statistics, such as, *Colless' index* (Colless, 1982), *Sackin's index* (Sackin, 1975; Tajima, 1983), *number of cherries* (McKenzie and Steel, 2000), *Aldous' shape statistic sequence* (Aldous, 2001) and *runs statistics* (Ford et al, 2009), can be obtained efficiently from appropriate coarsenings of the *vintaged and shaped* n-coalescent $\{G^{\uparrow}(t)\}$.

1.2 Statistical Motivation

The n-coalescent provides the basic probability model underlying statistical experiments of interest in population genetics. It arises as a prior mixture over $\mathscr{C}_n \otimes \mathbb{R}_+^n$, the set of all binary coalescent trees with branch lengths (recall that $\mathscr{C}_n$ was defined as

the set of sequences of length n with values in $\mathbb{C}_n$):

$$\mathscr{C}_n \otimes \mathbb{R}_+^n := \left\{ c \otimes t := \left((c_n, t_n), (c_{n-1}, t_{n-1}), \ldots, (c_1, t_1) \right) : c \in \mathscr{C}_n, t \in \mathbb{R}_+^n \right\},$$

where c_i gives the partition of the sample into groups of individuals having reached a common ancestor when the sample has exactly i ancestors (hence, $c_n = \{\{1\}, \ldots, \{n\}\}$ is the initial condition, and c_i arises after the $n-i$-th merger/step of the coalescent), and t_i gives the amount of time during which the continuous-time coalescent remains at the value c_i. In other words, t_i gives the amount of time in the past during which the sample has exactly i extant ancestors. By convention, we declare that $t_1 = 0$. Figure 4 shows a coalescent tree for a sample of four individuals.

It is over this *hidden genealogy space* that one needs to integrate in order to obtain the likelihood of a parameter $\phi \in \boldsymbol{\Phi}$ on the basis of some observed data x_{obs}:

$$\begin{aligned} P(x_{obs}|\phi) &= \int_{\mathscr{C}_n \otimes \mathbb{R}_+^n} P(x_{obs} \,|\, c \otimes t, \phi)\, dP(c \otimes t \,|\, \phi) \\ &= \int_{\mathscr{C}_n \otimes \mathbb{R}_+^n} P(x_{obs}|c \otimes t, \phi) p_\phi(c \otimes t)\, d\mathbb{P}(c \otimes t), \end{aligned} \tag{1.1}$$

where in the last line we assume that the n-coalescent induced, ϕ-specific prior law $P(\cdot\,|\phi)$ on $\mathscr{C}_n \otimes \mathbb{R}_+^n$ is absolutely continuous with respect to some reference probability measure $\mathbb{P}$, and p_ϕ denotes the corresponding density.

Computational feasibility of "full-likelihood" methods that conduct Monte Carlo integration over the n-coalescent trees, in order to compute the likelihood of the observed data via importance samplers (e.g. Bahlo and Griffiths (1996); Birkner and Blath (2008); Griffiths and Tavare (1994, 1996); Iorio and Griffiths (2004); Slatkin (2002); Stephens and Donnelly (2000)), scales poorly with the size of the data and the complexity of the models in modern population genomics. Typical data sets contain DNA sequences of large homologous tracks of the genome for thousands of individuals in a population.

Given the massive scale of current genomic data, computational biologists are using "summary statistics" of the available data to reduce the computational burden of the inference procedure and make it "likelihood-free" on the basis of simulations from the finest genealogical resolutions (e.g. Beaumont et al (2009, 2002); Leuenberger and Wegmann (2009); Marjoram et al (2003); Pritchard et al (1999); Sisson et al (2007); Weiss and von Haeseler (1998)). For a survey of ABC methods in a more general setting see Marin et al (2012) and the references therein. However, these *approximate Bayesian computations* or ABC do not take advantage of the appropriate and sufficient coarsening (or *Markov lumping*) of the hidden genealogy space for the summary statistics being used.

Finding the appropriate resolution can be powerful in inference if the observed statistic of interest only depends on the original chain through this lumping. This can reduce large summations over excessively fine state spaces as noted in (Kemeny and Snell, 1960, p. 124). The Markov lumpings of the *vintaged and labeled* n-coalescent, $\{B^{\uparrow}(t)\}$, developed here can facilitate a computationally efficient and statistically sufficient approach to population-genetic inference based on the exact likelihood of

various families of population-genetic summary statistics. Such a sophisticated approach to inference based on summary statistics amounts to *approximate Bayesian computation done exactly* or ABCDE Sainudiin et al (2011).

Briefly, the sufficiency of the *unvintaged and sized* n-coalescent for the likelihood of a popular statistic called the *site frequency spectrum* or SFS is exploited by Sainudiin et al (2011) to conduct ABCDE. Such inference methods based on SFS and its linear combinations, including, the *number of segregating sites* (Watterson, 1975), *pairwise heterozygosity*, and *Tajima's D* (Tajima, 1989), are possible due to the Markov lumping $\mathscr{F} : \mathbb{C}_n \to \mathbb{F}_n$ that facilitates efficient integration over f-sequences or sequential realisations of $\{F^{\uparrow}(t)\}$ in order to compute the likelihood of the SFS, as opposed to the more conventional approach of integrating over (in importance sampling) or simulating from (in ABC) the unnecessarily finer resolution of c-sequences or sequential realisations of $\{C^{\uparrow}(t)\}$. Importance sampling using a controlled Markov chain is developed by augmenting the forward-time *unvintaged and sized* n-coalescent $\{F^{\downarrow}(t)\}$ in order to produce f-sequences that are conditioned on the observed SFS by Sainudiin et al (2011). Such inference algorithms are publicly available from `www.math.canterbury.ac.nz/~r.sainudiin/codes/lce/` under the terms of the GNU General Public License.

Inferential methods that are similar to Sainudiin et al (2011) but based on summary statistics that depend on one of the other coarsenings of the n-coalescents can be developed from the coalescent probabilities obtained in this paper. Here, we formally describe lumped Markov processes at more resolutions of the hidden genealogy space and point out classical summary statistics from phylogenetic and population genetic literature that can be sufficiently described by appropriate Markov lumpings. The backward-transition, state-specific, forward-transition and sequence-specific probabilities at each of our coalescent resolutions described in this paper constitute the applied probabilistic core of computationally efficient Monte Carlo algorithms for statistical inference in population genetics that can exploit the Markov lumping relations among the different coalescent resolutions. These formulaic descriptions of the probability structures, especially at the coarser resolutions, are a prerequisite for subsequent computationally efficient and exactly approximate inference such as ABCDE in the spirit of Sainudiin et al (2011) but on the basis of other appropriate summary statistics. We leave the inference algorithms that can build upon the probabilistic formulae developed here for future research.

In his 2012 plenary address to the International Society for Bayesian Analysis, Christian Robert said the following about the importance of lumping in simulation-intensive inference:

> Noise created by the simulation of pseudo-data is killing by orders of magnitude the information contained in the whole data. Therefore, it makes much more sense to first project in a smaller space, accepting that we are loosing information. But then because we are in a much smaller space epsilon (the approximation error in ABC algorithms) will be close to zero.

This paper provides the first steps towards such a projection into a smaller space and addresses the following unsolved issue with ABC methods mentioned by Marin et al (2012, p. 1179):

> The (ABC) method necessarily faces limitations imposed by large datasets or complex models, in that simulating pseudo-data may itself become an impossible task. Dimension-reducing technique that would simulate directly the summary statistics will quickly become necessary.

Finally, as early as 1960, Kemeny and Snell (1960, p. 124) observe the following about a lumped process:

> It is also often the case in applications that we are only interested in questions which relate to this coarser analysis of the possibilities. Thus it is important to be able to determine whether the new process can be treated by Markov chain methods.

It is exactly this observation about a lumped Markov process in the coalescent context that led to this paper. Furthermore, the transition probabilities and Markov lumpings we develop here allow us to consistently move between different n-coalescent resolutions – a crucial strategy that could be exploited in simulation-intensive inference methods such as ABC. In this paper we take the necessary applied probabilistic steps towards realizing the potential for computationally efficient and statistically sufficient inference from population genetic statistics of today's massive genomic data.

1.3 Outline

The rest of this paper is organized as follows. In Sect. 2, we introduce the main notation and definitions that we will use later. In Sect. 3, we describe and study the six coalescent resolutions described in the Introduction. In Sect. 4, we provide concrete justifications of the interest of Markov lumpings. In particular, we give several examples of tree shape statistics depending only on a coarser resolution of the "full" Kingman coalescent. We also give an example of inference based on the coalescent with epoch times based on the observed *site frequency spectrum*. Finally, all the proofs are given in the Appendix (Sect. 5).

2 Preliminaries

Let $\mathbb{N} := \{1,2,3,\ldots\}$ denote the set of natural numbers. Let $\mathbb{Z}_+ := \{0,1,2,\ldots\}$ and $\mathbb{Z}_- := \{0,-1,-2,\ldots\}$ denote the set of non-negative and non-positive integers, respectively. For any set A, let $|A|$ denote its cardinality or the number of elements in it. Let $[n:n']_- := \{n,n-1,\ldots,n'+1,n'\}$ denote the linearly ordered descending index set from n to $n' \leq n$, where $n,n' \in \mathbb{Z}$ and let $[n]_- := [n:1]_- = \{n,n-1,\ldots,2,1\}$. Similarly, let $[n':n]_+ := \{n',n'+1,\ldots,n-1,n\}$ denote the linearly ordered ascending index set from n' to $n \geq n'$, where $n,n' \in \mathbb{Z}$ and let $[n]_+ := [1:n]_+ = \{1,2,\ldots,n-1,n\}$.

In what follows, we will consider time in two directions. A process with exponent $\uparrow$ will run *backwards-in-time*, that is the arrow of time will point towards the past. The basic example of this is Kingman's n-coalescent $\{C^{\uparrow}(t)\}$. On the other hand, a process with exponent $\downarrow$ will run *forwards-in-time*, that is from the epoch at which there is only one ancestor to the whole sample, until the present. The fact that backwards

processes are indexed by $[n]_-$ and forwards processes by $[n]_+$ guarantees that for any $i \in \{1,\ldots,n\}$, the value at step i of both chains correspond to the period of time during which the sample has exactly i ancestors.

The n-coalescent resolutions (with exception of the unvintaged and sized n-coalescent and the lineage death process) induce well-known types of phylogenetic trees on n leaves. We will formally define the trees we will observe throughout this paper.

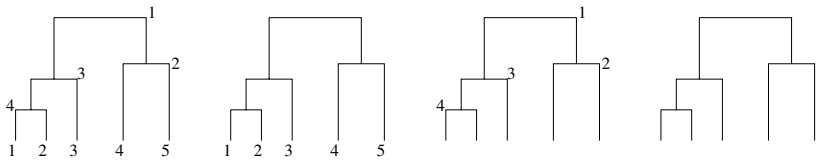


Fig. 2 Example for a ranked labeled tree with leaf label set $\mathfrak{L} = \{1,2,3,4,5\}$, a labeled tree with $\mathfrak{L} = \{1,2,3,4,5\}$, a ranked tree shape and a tree shape (from left to right).

Definition 1 We define the following trees as in Semple and Steel (2003, Sect. 2.4).

(*i*) A *ranked labeled tree* on n leaves is a rooted binary tree with unique leaf labels from the label set $\mathfrak{L}$. The interior vertices have a total order $<$ assigned, which satisfies the following requirements. The root is the minimum in this order, and if v is an interior vertex v which is on the path from an interior vertex w to a leaf, then $w < v$. Then, the root of the tree is given rank 1, the second smallest element in this total order has rank 2, etc.

(*ii*) A *labeled tree* on n leaves is a ranked labeled tree where the total order with the ranks are omitted.

(*iii*) A *ranked tree shape* on n leaves is a ranked labeled tree where the leaf labels are omitted.

(*iv*) A *tree shape* on n leaves is a labeled tree where the leaf labels are omitted.

See Fig. 2 for an example.

3 Six coalescent resolutions

Let us now describe the six resolutions of the Kingman-Tajima coalescent on which we will concentrate in the rest of this paper. One may observe that the forward-in-time and backward-in-time transition probabilities are well-defined for all these genealogical processes, because there is a unique initial state for each such process (so that the probability of visiting a given state is well-defined and we can use Bayes' formula to find the forward-in-time transition probabilities).

3.1 The block number resolution

This is the coarsest and by far the simplest resolution of the coalescent. Indeed, since we assume that the coalescent starts from n blocks and that only a single pairwise

merger can occur at any step, the Markov chain $\{H^{\uparrow}(t), t \in [n]_{-}\}$ is almost surely equal to the sequence $\{n, n-1, \ldots, 1\}$. In other words, the state-space $\mathscr{H}_n$ of all possible h-sequences is reduced to the single point $\{n, n-1, \ldots, 1\}$ and for every i,

$$P(H^{\uparrow}(i) = i) = 1.$$

Of course, the same holds for the forward chain $\{H^{\downarrow}(t)\}$ (recall from Sect. 2 that *backward* chains are indexed by $[n]_{-}$ and *forward* chains by $[n]_{+}$), that is,

$$P(H^{\downarrow}(i) = i) = 1.$$

3.2 The vintaged and labeled resolution

At this finest resolution in this study, in each epoch, we keep track of the blocks formed by the labels of the individuals of our sample having reached a common ancestor before or at this step, as well as of the *epoch* at which each of these ancestral blocks was created as we follow the genealogy of our sample back through time. This Markov process can be lumped to any other process we will introduce below.

First we derive the state space $\mathbb{B}_n$. Recall that $\mathbb{C}_n$ denotes the set of all set partitions of the label set $\mathfrak{L} = \{1, 2, \ldots, n\}$ of n-samples. Let $|c_a|$ denote the number of elements in $c_a \in \mathbb{C}_n$. Denote the set of all set partitions with i blocks by $\mathbb{C}_n^{(i)}$, so that $\mathbb{C}_n = \bigcup_{i=1}^{n} \mathbb{C}_n^{(i)}$. Let $c_i := \{c_{i,1}, c_{i,2}, \ldots, c_{i,i}\} \in \mathbb{C}_n^{(i)}$ denote the i elements of c_i. The partial ordering $\overset{\cdot\cdot\cdot}{\prec}_c$ on $\mathbb{C}_n$ is based on the immediate precedence relation $\prec_c$:

$$c_{i'} \prec_c c_i \Leftrightarrow c_{i'} = \big(c_i \setminus \{c_{i,j}, c_{i,k}\}\big) \cup (c_{i,j} \cup c_{i,k}) \text{ for some } j \neq k, j, k \in \{1, \ldots, i\}.$$

In words, $c_{i'} \prec_c c_i$, read as $c_{i'}$ immediately precedes c_i, means that $c_{i'}$ can be obtained from c_i by coalescing a distinct pair of elements in c_i. Thus, $c_{i'} \prec_c c_i$ implies $|c_{i'}| = |c_i| - 1$.

Let the coalescent epochs be labeled $n, n-1, \ldots, 1$ as we go back in time, with epoch n starting at the present, and a new epoch starting with each coalescent event. Thus, as we have already seen with the lineage death process, there are k lineages during epoch k. We say that a lineage identified by $c_{i,j}$ in the i-th epoch, i.e. the lineage that subtends the sample labels in the set $c_{i,j}$, is of $m_{i,j}$ vintage if $c_{i,j}$ originated at the start of epoch $m_{i,j}$ (going back in time). We also say that $m_{i,j}$ is the coalescent-epoch vintage or simply the vintage of $c_{i,j}$. We notate such lineage-vintage pairs, $\texttt{lineage}^{\langle \texttt{vintage} \rangle}$, or *vintaged lineages* by $b_{i,j} := c_{i,j}^{\langle m_{i,j} \rangle}$ and we let

$$b_i := \{b_{i,1}, b_{i,2}, \ldots, b_{i,i}\} := \left\{c_{i,1}^{\langle m_{i,1} \rangle}, c_{i,2}^{\langle m_{i,2} \rangle}, \ldots, c_{i,i}^{\langle m_{i,i} \rangle}\right\}$$

denote the i vintaged lineages in epoch i formed by pairing each element $c_{i,j}$ of c_i with its respective vintage $m_{i,j} \in \{n, n-1, \ldots, i\}$. Let the set of such b_i's be $\mathbb{B}_n^{(i)}$ and let $\mathbb{B}_n := \bigcup_{i=1}^{n} \mathbb{B}_n^{(i)}$. Thus, $\mathbb{B}_n$ is a vintage augmentation of $\mathbb{C}_n$. We say that $b_{i'}$ immediately precedes b_i and write $b_{i'} \prec_b b_i$, if and only if:

$$b_{i'} = b_i \setminus \{c_{i,j}^{\langle m_{i,j} \rangle}, c_{i,k}^{\langle m_{i,k} \rangle}\} \bigcup (c_{i,j} \cup c_{i,k})^{\langle |b_i| - 1 \rangle}, \text{ for some } j \neq k \in \{1, 2, \ldots, |b_i|\}.$$

In words, $b_{i'} \prec_b b_i$ means that $b_{i'}$ can be obtained from b_i by coalescing any distinct pair of lineages in b_i and by updating the coalesced lineage's vintage tag to that of the new epoch label. Let $b := (b_n, b_{n-1}, \ldots, b_1)$ be a sequence of states in $\mathbb{B}_n$ that consecutively satisfy the immediate precedence relation $\prec_b$ and let $\mathscr{B}_n$ be the set of such b-sequences. A b-sequence for $n = 4$ is given in Table 1.

Proposition 1 (Probabilities over $\mathbb{B}_n$) *The backward transition probabilities $P(b_{i-1}|b_i)$ of the Markov chain $\{B^{\uparrow}(k)\}_{k\in[n]_-}$ on $\mathbb{B}_n$ with initial state $b_n = \{\{1\}^{\langle n\rangle}, \{2\}^{\langle n\rangle}, \ldots, \{n\}^{\langle n\rangle}\}$ and final absorbing state $b_1 = \{\{1,2,\ldots,n\}^{\langle 1\rangle}\}$, the probability $P(b_i)$ of visiting a state b_i, the forward transition probabilities $P(b_i|b_{i-1})$ of the Markov chain $\{B^{\downarrow}(k)\}_{k\in[n]_+}$ on $\mathbb{B}_n$ with initial state b_1 and final absorbing state b_n, and the probability $P(b)$ of a b-sequence in $\mathscr{B}_n$ are:*

$$P(b_{i-1}|b_i) = \begin{cases} \binom{i}{2}^{-1} & \text{if } b_{i-1} \prec_b b_i, b_i \in \mathbb{B}_n^{(i)} \\ 0 & \text{otherwise} \end{cases}, \tag{3.1}$$

$$P(b_i) = \frac{\prod_{j=1}^{i'} |c_{i,j}|!(|c_{i,j}|-1)(|c_{i,1:j}|-j-1-m_{i,j}+i)!}{n!(n-1)!\,(i!(i-1)!)^{-1}\prod_{j=1}^{i'-1}(|c_{i,1:j}|-j-m_{i,j+1}+i)!}, \tag{3.2}$$

$$P(b_i|b_{i-1}) = \frac{P(b_i)P(b_{i-1}|b_i)}{P(b_{i-1})}, \tag{3.3}$$

$$P(b) = P(b_{n-1}|b_n)P(b_{n-2}|b_{n-1})\cdots P(b_1|b_2) = \frac{2^{n-1}}{n!\,(n-1)!} = \frac{1}{|\mathscr{B}_n|}, \tag{3.4}$$

where if we order the blocks of b_i in such a way that $m_{i,1} \le m_{i,2} \le \cdots \le m_{i,i}$, then $c_{i,1:j} := c_{i,1} \cup \cdots \cup c_{i,j}$ and $i' := \max\{j : m_{i,j} < n\}$.

Proof See Proof 1 in the Appendix.

Proposition 2 (Bijection between ranked labeled trees and b-sequences) *There is a bijection between the set of ranked labeled trees on n leaves and $\mathscr{B}_n$, the set of b-sequences.*

The proof of Prop. 2 is straightforward and is therefore omitted.

3.3 Unvintaged and labeled resolution

Kingman's *unvintaged and labeled* Markov chain over $\mathbb{C}_n$ can be obtained as a Markov lumping of the vintaged and labeled chain $\{B^{\uparrow}(k)\}_{k\in[n]_-}$, by omitting the epoch vintages from the states in $\mathbb{B}_n$. Let $c := (c_n, c_{n-1}, \ldots, c_1)$ be a c-sequence or coalescent sequence obtained from the sequence of states visited by a sequential realization of the backward in time Markov chain $\{C^{\uparrow}(k)\}_{k\in[n]_-}$, and recall that $\mathscr{C}_n$ denotes the set of such c-sequences:

$$\mathscr{C}_n := \{c := (c_n, c_{n-1}, \ldots, c_1) : c_i \in \mathbb{C}_n^i, c_{i-1} \prec_c c_i, i \in \{n, n-1, \ldots, 2\}\}$$

A c-sequence for $n = 4$ is given in Table 1.

Proposition 3 (Probabilities over $\mathbb{C}_n$) *The backward transition probabilities $P(c_{i-1}|c_i)$ of the Markov chain $\{C^{\uparrow}(k)\}_{k\in[n]_-}$ on $\mathbb{C}_n$ with initial state $c_n = \{\{1\},\ldots,\{n\}\}$ and final absorbing state $c_1 = \{\{1,\ldots,n\}\}$, the probability $P(c_i)$ of visiting a state c_i, the forward transition probabilities $P(c_i|c_{i-1})$ of the Markov chain $\{C^{\downarrow}(k)\}_{k\in[n]_-}$ on $\mathbb{C}_n$ with initial state c_1 and final absorbing state c_n, and the probability $P(c)$ of a c-sequence are:*

$$P(c_{i-1}|c_i) = \begin{cases} \binom{i}{2}^{-1} & : \text{if } c_{i-1} \prec_c c_i,\ c_i \in \mathbb{C}_n^{(i)}, \\ 0 & : \text{otherwise,} \end{cases} \tag{3.5}$$

$$P(c_i) = \frac{(n-i)!\,i!\,(i-1)!}{n!\,(n-1)!} \prod_{j=1}^{i} |c_{i,j}|!, \tag{3.6}$$

$$P(c_i|c_{i-1}) = \frac{2}{(n-i+1)\binom{|c_{i,j_*}|+|c_{i,j'_*}|}{|c_{i,j_*}|}}, \tag{3.7}$$

$$P(c) = \prod_{i=2}^{n} P(c_{i-1}|c_i) = \frac{2^{n-1}}{n!\,(n-1)!} = \frac{1}{|\mathscr{C}_n|}, \tag{3.8}$$

where c_{i,j_*} and c_{i,j'_*} are the two blocks created by the split of a block of c_{i-1} that gives rise to c_i in the forward chain $\{C^{\downarrow}(t)\}$.

Proof Equations (3.5) and (3.6) are established in (Kingman, 1982a, (2.2)) and (Kingman, 1982a, (2.3)), respectively, and Eq. (3.8) follows from the Markov property of $\{C^{\uparrow}(t)\}$ and Eq. (3.5). See Proof 2 for proofs of Eq. (3.7) and of Eq. (3.6) from Eq. (3.2).

Proposition 4 (Bijection between ranked labeled trees and c-sequences) *There is a bijection between ranked labeled trees on n leaves and $\mathscr{C}_n$, the set of c-sequences.*

Again, the proof of this straightforward result is omitted.

Observe that the number of elements in $\mathbb{C}_n$ is the number of set partitions of a set of size n which is $\text{Bell}(n)$, the n-th Bell number

$$|\mathbb{C}_n| = \text{Bell}(n) := \sum_{j=0}^{n} S_n^{(j)}, \quad . \tag{3.9}$$

where $S_n^{(j)}$ is the Stirling number of the second kind with parameters n and j.

Remark 1 Our forward-in-time Markov chain $\{C^{\downarrow}(k)\}_{k\in[n]_+}$ on $\mathbb{C}_n$ is different from Aldous' *beta-splitting model* (Aldous, 2001). The beta-splitting model also produces bipartitions of a label set forward in time as a *Markov branching* model. The distinguishing feature of the beta-splitting model is its recursive repetition of the same bipartitioning or splitting process anew on elements of a partition of the label set. Therefore the beta-splitting model only induces labeled trees, but no ranking. When the parameter $\beta = 0$, the beta-splitting model induces the same distribution on labeled trees (without ranking) as the vintaged/unvintaged and labeled n-coalescent. In Sect. 4.2 we revisit Aldous' shape statistics that originated under the beta-splitting model from the lumped Markov chains of Sect. 3.6, $\{F^{\uparrow}(k)\}_{k\in[n]_+}$ and $\{F^{\downarrow}(k)\}_{k\in[n]_+}$, on $\mathbb{F}_n$.

3.4 Vintaged and sized resolution

Here we keep track of the sizes of the blocks (i.e., the number of descendants of each ancestor) along with their vintages. Consider the coalescent epoch i during which there are i blocks. Let $d_{i,j}$ denote the size of a block with coalescent vintage $j \in \{1,2,\ldots,n-1\}$ during the i-th epoch (recall that there are at most one such block). By convention, $d_{i,j} = 0$ if there are no such block. Then, $d_{i,n} = n - \sum_{j=1}^{n-1} d_{i,j}$ is the number of singleton blocks (or leaf lineages) during the i-th epoch.

Let us now represent the state of the coalescent during the i-th epoch by the vector $d_i := (d_{i,1}, d_{i,2}, \ldots, d_{i,n-1})$. The state space of such *vintaged* and *sized* ancestral partition during the i-th epoch can be defined by the set

$$\mathbb{D}_n^{(i)} := \left\{ d_i \in \mathbb{Z}_+^{n-1} : \begin{cases} \sum_{j=1}^{i-1} d_{i,j} = 0, \\ \sum_{j=1}^{n-1} \mathbb{1}_{\mathbb{N}}(d_{i,j}) + \left(n - \sum_{j=1}^{n-1} d_{i,j}\right) = i, \\ d_{i,1} \neq 1, d_{i,2} \neq 1, \ldots, d_{i,n-1} \neq 1 \end{cases} \right\},$$

with $\mathbb{D}_n := \bigcup_{i=1}^{n} \mathbb{D}_n^{(i)}$.

Let e_i be the i-th unit vector of length $n-1$. We say that $d_{i'} \prec_d d_i \in \mathbb{D}_n^{(i)}$ if and only if:

$$d_{i'} = \begin{cases} d_i + (d_{i,j} + d_{i,k}) e_{i-1} & \\ \quad -d_{i,j} e_j - d_{i,k} e_k & \text{for some } i \leq j < k < n \text{ s.t. } d_{i,j} \neq 0, d_{i,k} \neq 0, \text{ or} \\ d_i + (d_{i,j} + 1) e_{i-1} - d_{i,j} e_j & \text{for some } i \leq j < n \text{ s.t. } d_{i,j} \neq 0, \text{ if } d_{i,n} \geq 1, \text{ or} \\ d_i + 2 e_{i-1} & \text{if } d_{i,n} \geq 2. \end{cases}.$$

A *d-sequence* $d := (d_n, d_{n-1}, \ldots, d_1)$ is obtained from a sequence of immediately preceding states in $\mathbb{D}_n$. Let $\mathscr{D}_n$ be the set of such d-sequences. A d-sequence for $n = 4$ is given in Table 1.

For every $d \in \mathscr{D}_n$, let $\beth(d)$ be the number of cherries in d, i.e. the number of times that we have $d_{i,n} - d_{i-1,n} = 2$ (corresponding to the merger of two singleton blocks) as i varies from n to 2. More formally,

$$\beth(d) := \sum_{i=2}^{n} \mathbb{1}_{\{2\}}(d_{i,n} - d_{i-1,n}).$$

Further, define $d_{i,1:j} := \sum_{k=1}^{j} d_{i,k}$, $m'_{i,j} := \min\{k > j : d_{i,k} > 0\}$ and $k_{i,j} := |\{m \leq j : d_{i,m} > 0\}|$.

Proposition 5 (Probabilities over $\mathbb{D}_n$) *The backward transition probabilities $P(d_{i-1}|d_i)$ of the Markov chain $\{D^{\uparrow}(k)\}_{k \in [n]_-}$ on $\mathbb{D}_n$ with initial state $d_n = (0,0,\ldots,0)$ and final absorbing state $d_1 = (n,0,0,\ldots,0)$, the probability $P(d_i)$ of visiting a state d_i, the forward transition probabilities $P(d_i|d_{i-1})$ of the Markov chain $\{D^{\downarrow}(k)\}_{k \in [n]_+}$ with initial state $d_1 = (n,0,0,\ldots,0)$ and final absorbing state $d_n = (0,0,\ldots,0)$, and the*

probability $P(d)$ of a a given d-sequence d are:

$$P(d_{i-1}|d_i) = \begin{cases} \binom{d_{i,n}}{d_{i,n}-d_{i-1,n}} \binom{i}{2}^{-1} & \text{if} \quad d_{i-1} \prec_d d_i \in \mathbb{D}_n^{(i)}, \\ 0 & \text{otherwise}. \end{cases} \tag{3.10}$$

$$P(d_i) = \frac{i!(i-1)!}{(n-1)!} \left(\frac{\prod_{j=1,d_{i,j}>0}^{n-1} (d_{i,j}-1)(d_{i,1:j}-k_{i,j}-j-1+i)!}{\prod_{j=1,d_{i,j}>0}^{n-1} (d_{i,1:j}-k_{i,j}-m'_{i,j}+i)!} \right), \tag{3.11}$$

$$P(d_i|d_{i-1}) = P(d_{i-1}|d_i) \frac{P(d_i)}{P(d_{i-1})}, \tag{3.12}$$

$$P(d) = \frac{2^{n-\beth(d)-1}}{(n-1)!}. \tag{3.13}$$

Proof See Proof 3 in the Appendix.

Observe that the chain $\{D^{\uparrow}(t)\}$ is a Markov lumping of the chain $\{B^{\uparrow}(t)\}$ based on the map $\mathscr{D}$ defined as follows: for every $b_k \in \mathbb{B}_n^{(k)}$,

$$\begin{aligned} \mathscr{D}(b_k) &= \mathscr{D}\left(\left\{ c_{k,1}^{\langle m_{k,1} \rangle}, \ldots, c_{k,k}^{\langle m_{k,k} \rangle} \right\} \right) \\ &:= \left(\sum_{j=1}^{k} |c_{k,j}| \, \mathbb{1}_{\{1\}}(m_{k,j}), \ldots, \sum_{j=1}^{k} |c_{k,j}| \, \mathbb{1}_{\{n-1\}}(m_{k,j}) \right). \end{aligned}$$

Proposition 6 (Bijection between ranked tree shapes and d-sequences) *There is a bijection between ranked tree shapes on n leaves and $\mathscr{D}_n$, the set of d-sequences.*

3.5 Vintaged and shaped resolution

Here we only track the vintages and forget the description of the blocks (or block sizes). Consider the coalescent epoch i during which there are i blocks. Let $g_{i,j}$ denote the presence ($g_{i,j} = 1$) or absence ($g_{i,j} = 0$) of a block with coalescent vintage $j \in \{1, 2, \ldots, n-1\}$ during the i-th epoch. The set of all vintaged and shaped coalescent states during the i-th epoch is thus defined by

$$\mathbb{G}_n^{(i)} := \left\{ g_i \in \{0,1\}^{n-1} : g_{i,i} = 1, \sum_{j=1}^{i-1} g_{i,j} = 0, \sum_{j=1}^{n-1} g_{i,j} \leq i \right\},$$

and we set $\mathbb{G}_n := \bigcup_{i=1}^{n} \mathbb{G}_n^{(i)}$. In this description of $\mathbb{G}_n^{(i)}$, the $g_{i,i} = 1$ represents the block that just arose at the beginning of the i-th epoch. Of course no blocks can carry a vintage smaller than the current epoch i, and furthermore the number of blocks with vintages smaller than n cannot exceed the total number i of blocks. Finally, all blocks having vintage n are singleton blocks that have not yet coalesced, and so

$$g_{i,n} = i - \sum_{j=1}^{n-1} g_{i,j}.$$

We say that $g_{i'} \prec_g g_i \in \mathbb{G}_n^{(i)}$ if and only if:

$$g_{i'} = \begin{cases} g_i + e_{i-1} - e_j - e_k & \text{for some } i \le j < k < n \text{ s.t. } g_{i,j} = g_{i,k} = 1, \text{ or} \\ g_i + e_{i-1} - e_j & \text{for some } i \le j < n \text{ s.t. } g_{i,j} = 1, \text{ if } g_{i,n} \ge 1, \text{ or} \\ g_i + e_{i-1} & \text{if } g_{i,n} \ge 2. \end{cases}$$

A *g-sequence* $g := (g_n, g_{n-1}, \ldots, g_1)$ is obtained from a sequence of immediately preceding states in $\mathbb{G}_n$. Let $\mathscr{G}_n$ be the set of such g-sequences. A g-sequence for $n = 4$ is given in Table 1.

Let $\beth(g) := \sum_{i=2}^{n} \mathbb{1}_{\{2\}}(g_{i,n} - g_{i-1,n})$ be the number of cherries in g, and let $\mathscr{G} : \mathbb{D}_n \to \mathbb{G}_n$ be the size-dropping map:

$$\mathscr{G}(d_k) = \mathscr{G}((d_{k,1}, \ldots, d_{k,n})) := (\mathbb{1}_{\mathbb{N}}(d_{k,1}), \ldots, \mathbb{1}_{\mathbb{N}}(d_{k,n-1})) = (g_{k,1}, \ldots, g_{k,n-1}). \tag{3.14}$$

This map induces a Markov lumping of the chain $\{D^{\uparrow}(t)\}$ into the chain $\{G^{\uparrow}(t)\}$ described in the following Proposition.

Proposition 7 (Probabilities over $\mathbb{G}_n$) *The backward transition probabilities $P(g_{i-1}|g_i)$ of the Markov chain $\{G^{\uparrow}(k)\}_{k\in[n]_-}$ on $\mathbb{G}_n$ with initial state $g_n = (0,0,\ldots,0) \in \mathbb{G}_n^{(n)}$ and final absorbing state $g_1 = (1,0,0,\ldots,0) \in \mathbb{G}_n^{(1)}$, the probability $P(g_i)$ of visiting a state g_i, the forward transition probabilities $P(g_i|g_{i-1})$ of the Markov chain $\{G^{\downarrow}(k)\}_{k\in[n]_+}$ on $\mathbb{G}_n$ with initial state $g_1 = (1,0,0,\ldots,0)$ and final absorbing state $g_n = (0,0,\ldots,0)$, and the probability $P(g)$ of a g-sequence g are:*

$$P(g_{i-1}|g_i) = \begin{cases} \binom{g_{i,n}}{g_{i,n}-g_{i-1,n}} \binom{i}{2}^{-1} & \text{if} \quad g_{i-1} \prec_g g_i \in \mathbb{G}_n^{(i)}, \\ 0 & \text{otherwise}, \end{cases} \tag{3.15}$$

$$P(g_i) = P(\mathscr{G}^{-1}(g_i)) = \sum_{d_j \in \mathscr{G}^{-1}(g_i)} P(d_j), \tag{3.16}$$

$$P(g_i|g_{i-1}) = P(g_{i-1}|g_i)\frac{P(g_i)}{P(g_{i-1})}, \tag{3.17}$$

$$P(g) = \prod_{i=2}^{n} P(g_{i-1}|g_i) = \frac{2^{n-\beth(g)-1}}{(n-1)!}. \tag{3.18}$$

Proof See Proof 4 in the Appendix.

Remark 2 The space of g-sequences is in bijection with Tajima's evolutionary relationships (Tajima, 1983, Figures 1-4) and this is why we refer to the Markov chain $\{G^{\uparrow}(k)\}_{k\in[n]_-}$ on $\mathbb{G}_n$ as Tajima's n-coalescent. This is the first Markov description of Tajima's evolutionary relationships in the coalescent framework of Kingman.

Proposition 8 *The number of elements in $\mathbb{G}_n$ is*

$$|\mathbb{G}_n| = \text{Fibo}(n+1) \ , \tag{3.19}$$

where Fibo(n) *is the n-th Fibonacci number.*

Proof See *Proof 5 in the Appendix.*

Proposition 9 (Bijection between ranked tree shapes and g-sequences) *There is a bijection between the set of ranked tree shapes on n leaves and $\mathscr{G}_n$, the set of g-sequences.*

Proposition 10 (The ranked labeled trees of an g-sequence) *Let $g \in \mathscr{G}_n$ be any given g-sequence and let $b \in (\mathscr{G} \circ \mathscr{D})^{-1}(g) := \{\mathscr{D}^{-1}(d) : d \in \mathscr{G}^{-1}(g)\}$ be a corresponding b-sequence. Then the number of b-sequences (which is the number of ranked labeled trees) corresponding to the given g is*

$$|(\mathscr{G} \circ \mathscr{D})^{-1}(g)| = 2^{1-n} n!\,(n-1)!\,P(g) = n!\,2^{-\beth(g)} \ , \tag{3.20}$$

where $\beth(g)$ is the number of cherries of the ranked tree shape induced by g. The conditional probability of b or c given g is

$$P(b|g) = P(c|g) = 2^{\beth(g)}/n! \ . \tag{3.21}$$

Proof See Proof 6 in the Appendix.

3.6 Unvintaged and sized resolution

Here we track the sizes of the blocks and disregard their labels and vintages. The unvintaged and sized resolution is mentioned as a lumped Markov chain of the unvintaged and labeled resolution and termed the 'label-killed' process by Kingman (1982b, 5.2). Tavaré (1983, p. 136-137) calls it the 'family-size process' as part of the nomenclature of a more general birth-death-immigration process of Kendall (1975). The transition probabilities of this Markov chain $\{F^{\uparrow}(k)\}_{k\in[n]_-}$ are not explicitly developed in Kingman (1982b) or Tavaré (1983). They have been developed in Saintudiin et al (2011) to resolve the hidden genealogy space just enough to prescribe the likelihood of a popular classical statistic called the site frequency spectrum and its linear combinations. We briefly retrace $\{F^{\uparrow}(k)\}_{k\in[n]_-}$ and its companion chains to show that they can provide the sampling distribution of a large family of shape statistics including several classical ones. The significantly smaller state space of $\{F^{\uparrow}(k)\}_{k\in[n]_-}$ allows for a computationally efficient and statistically sufficient inference based on these statistics.

Consider the coalescent epoch at which there are i lineages. Let $f_{i,j}$ denote the number of blocks of size j (or lineages ancestral to j individuals) at this epoch. Let us summarize these numbers by the vector $f_i := (f_{i,1}, f_{i,2}, \ldots, f_{i,n})$. Then, the space of f_i's is the set of integer partitions of n composed of i positive integers and is defined by

$$\mathbb{F}_n^{(i)} := \left\{ f_i := (f_{i,1}, f_{i,2}, \ldots, f_{i,n}) \in \mathbb{Z}_+^n : \sum_{j=1}^{n} j f_{i,j} = n, \sum_{j=1}^{n} f_{i,j} = i \right\}.$$

Let the set of such frequencies over all epochs be $\mathbb{F}_n := \bigcup_{i=1}^{n} \mathbb{F}_n^{(i)}$, the frequency representation of the integer partitions of n, i.e. the solutions to the Diophantine equation $\{(p_1, p_2, \ldots, p_n) \in \mathbb{Z}_+^n : \sum_{i=1}^{n} i p_i = n\}$. Thus, the cardinality of $\mathbb{F}_n$ is the number of integer partitions of n:

$$|\mathbb{F}_n| = 1 + \sum_{k=1}^{\lfloor n/2 \rfloor} \mathfrak{p}(k, n-k), \qquad \text{where}$$

$$\mathfrak{p}(k,n) = \begin{cases} 0 & \text{if} \quad k > n \\ 1 & \text{if} \quad k = n \\ \mathfrak{p}(k+1,n) + \mathfrak{p}(k,n-k) & \text{otherwise.} \end{cases} \tag{3.22}$$

Let us define an *f-sequence* f as follows:

$$f := (f_n, f_{n-1}, \ldots, f_1) \in \mathscr{F}_n := \left\{ f : f_i \in \mathbb{F}_n^{(i)}, f_{i-1} \prec_f f_i, \forall i \in \{2, \ldots, n\} \right\},$$

where $\prec_f$ is the immediate precedence relation defined by (here, e_j denotes the j-th unit vector of length n):

$$f_{i'} \prec_f f_i \Leftrightarrow f_{i'} = f_i - e_j - e_k + e_{j+k} \quad \text{for some } 1 \leq j,k \leq n.$$

Thus, $\mathscr{F}_n$ is the set of f-sequences with n samples. One can see $\mathscr{F}_n$ as the set of the frequencies of the cardinalities of c-sequences in $\mathscr{C}_n$. Indeed, if we define the map $\mathscr{F} : \mathbb{C}_n \to \mathbb{F}_n$ by

$$\mathscr{F}(c_i) := \left(\sum_{h=1}^{i} \mathbb{1}_{\{1\}}(|c_{i,h}|), \ldots, \sum_{h=1}^{i} \mathbb{1}_{\{n\}}(|c_{i,h}|) \right), \tag{3.23}$$

then $\mathscr{F}$ induces a Markov lumping of the chain $\{C^{\uparrow}(t)\}$ into the chain $\{F^{\uparrow}(t)\}$.

An f-sequence f written as $(f_n, f_{n-1}, \ldots, f_1)$ is an $n \times n$ matrix:

$$f := \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix} := \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n-1} & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n-1} & f_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{n-1,1} & f_{n-1,2} & \cdots & f_{n-1,n-1} & f_{n-1,n} \\ f_{n,1} & f_{n,2} & \cdots & f_{n,n-1} & f_{n,n} \end{pmatrix}$$

Note that $\mathscr{F}_n$ indexes an equivalence class in $\mathscr{C}_n$ via the inverse map $\mathscr{F}^{-1}$.

Before detailing the transition probabilities of the unvintaged and sized coalescent, let us define a shape statistic triple of any $f \in \mathscr{F}_n$. Let us denote the entry-wise maximum or minimum of a vector x by $\max\langle x \rangle$ and $\min\langle x \rangle$, respectively. There are $n-1$ coalescence events in any f. Define $\beth(f)$ as the number of events resulting from the coalescence of a pair of singleton blocks (or leaves). As in the previous paragraphs, such an event is called a cherry. Next, define $\daleth(f)$ as the number of events that arise from coalescing two blocks of distinct sizes. Let the number of the remaining events in f be defined as $\widehat{\beth}(f)$. Thus, $\widehat{\beth}(f)$ is the number of events resulting from the coalescence of two blocks of equal size that are not cherries. A distinctly-sized

split of a block of size i gives rise to two blocks of size i_1 and i_2, such that $i_1 \neq i_2$ and $i = i_1 + i_2$. In formulae, the above is,

$$\beth(f) := \sum_{i=2}^{n} \mathbb{1}_{\{1\}}(f_{i-1,2} - f_{i,2}), \tag{3.24}$$

$$\daleth(f) := \sum_{i=2}^{n} \mathbb{1}_{\{1\}}(\max\langle f_i - f_{i-1}\rangle), \tag{3.25}$$

$$\widehat{\beth}(f) := n - 1 - \daleth(f) - \beth(f). \tag{3.26}$$

Denoting the entry-wise or Hadamard product by $\boxtimes$, let us define $\ddot{f}_i$ as the number of blocks having the same size as the block that was split at the beginning of the i-th epoch (forward in time) and the corresponding *split frequency vector* $\ddot{\Lambda}(f) = \ddot{f} := (\ddot{f}_2, \ddot{f}_3, \ldots, \ddot{f}_n)$. For a given f-sequence f, we have

$$\ddot{f}_i := f_{i-1,-\min\langle (f_i - f_{i-1}) \boxtimes (1,2,\ldots,n)\rangle}.$$

For example, if there were four blocks of size three each and one of these four blocks splits at the beginning of the i-th epoch, then $\ddot{f}_i = 4$.

Proposition 11 (Probabilities over $\mathbb{F}_n$) *The backward transition probabilities $P(f_{i-1}|f_i)$ of the Markov chain $\{F^{\uparrow}(k)\}_{k\in[n]_-}$ on $\mathbb{F}_n$ with initial state $f_n = (n,0,\ldots,0) \in \mathbb{F}_n^{(n)}$ and final absorbing state $f_1 = (0,0,\ldots,1) \in \mathbb{F}_n^{(1)}$, the probability $P(f_i)$ of visiting a state f_i, the forward transition probabilities $P(f_i|f_{i-1})$ of the Markov chain $\{F^{\downarrow}(k)\}_{k\in[n]_+}$ on $\mathbb{F}_n$ with initial state $f_1 = (0,0,\ldots,1)$ and final absorbing state $f_n = (n,0,\ldots,0)$, and the probability $P(f)$ of an f-sequence f are:*

$$P(f_{i-1}|f_i) = \begin{cases} f_{i,j} f_{i,k} \binom{i}{2}^{-1} & \text{if } f_{i-1} = f_i - e_j - e_k + e_{j+k},\ j \neq k \\ \binom{f_{i,j}}{2}\binom{i}{2}^{-1} & \text{if } f_{i-1} = f_i - 2e_j + e_{2j}, \\ 0 & \text{otherwise}, \end{cases} . \tag{3.27}$$

$$P(f_i) = \frac{i!}{\prod_{j=1}^{i} f_{i,j}!} \binom{n-1}{i-1}^{-1}, \tag{3.28}$$

$$P(f_i|f_{i-1}) = \begin{cases} 2f_{i-1,j+k}(n-i+1)^{-1} & \text{if } f_i = f_{i-1} + e_j + e_k - e_{j+k},\ j \neq k, \\ f_{i-1,2j}(n-i+1)^{-1} & \text{if } f_i = f_{i-1} + 2e_j - e_{2j},\ j = k, \\ 0 & \text{otherwise}, \end{cases} \tag{3.29}$$

$$P(f) = \frac{2^{\daleth(f)}}{(n-1)!} \prod_{i=2}^{n} \ddot{f}_i, \quad . \tag{3.30}$$

Proof See Proof 7 in the Appendix.

We have seen earlier in this section that $\{F^{\uparrow}(t)\}$ was a Markov lumping of $\{C^{\uparrow}(t)\}$ through the function $\mathscr{F}$ defined in Eq. (3.23). It can also be considered as a coarsening (or Markov lumping) of $\{D^{\uparrow}(t)\}$ via the following vintage-dropping map $\mathscr{F}' : \mathbb{D}_n \to \mathbb{F}_n$:

$$\mathscr{F}'(d_k) := \left(n - \sum_{j=1}^{n-1} d_{k,j}, \sum_{j=1}^{n-1} \mathbb{1}_{\{2\}}(d_{k,i}), \ldots, \sum_{j=1}^{n-1} \mathbb{1}_{\{n\}}(d_{k,i}) \right). \tag{3.31}$$

Proposition 12 (The ranked labeled trees of an f-sequence) *Let $f \in \mathscr{F}_n$ be any given f-sequence and let $\mathscr{F}^{-1}(f)$ be the set of all corresponding c-sequences. Then the cardinality of $\mathscr{F}^{-1}(f)$ (which is also the number of ranked labeled trees corresponding to the given f) is*

$$|\mathscr{F}^{-1}(f)| = 2^{1-n}\, n!\, (n-1)!\, P(f) = n!\, 2^{\urcorner(f)+1-n} \prod_{i=2}^{n} \ddot{f}_i, \tag{3.32}$$

and the conditional probability of $c \in \mathscr{F}^{-1}(f)$ given f is

$$P(c|f) = \frac{2^{-\urcorner(f)+n-1}}{n!} \prod_{i=2}^{n} \ddot{f}_i^{-1} = \frac{1}{|\mathscr{F}^{-1}(f)|}. \tag{3.33}$$

Proof See Proof 8 in the Appendix.

Proposition 13 (The ranked tree shapes of an f-sequence) *Let $f \in \mathscr{F}_n$ be any given f-sequence and let $d \in \mathscr{F}'^{-1}(f)$ and $g \in \mathscr{G}(\mathscr{F}'^{-1}(f)) := \{\mathscr{G}(d) : d \in \mathscr{F}'^{-1}(f)\}$ be a corresponding d- and g-sequence, respectively. The number of ranked tree shapes corresponding to the given f is*

$$|\mathscr{F}'^{-1}(f)| = |\mathscr{G}(\mathscr{F}'^{-1}(f))| = 2^{-\widehat{\beth}(f)} \prod_{i=2}^{n} \ddot{f}_i \ , \tag{3.34}$$

and the conditional probability of g given f is

$$P(g|f) = 2^{\widehat{\beth}(f)} \left(\prod_{i=2}^{n} \ddot{f}_i \right)^{-1} \ . \tag{3.35}$$

Proof See Proof 9 in the Appendix.

3.7 Examples

Next we provide some concrete examples of n-coalescent sequences at various resolutions for small n and calculate $P(f)$, $|\mathscr{F}^{-1}(f)|$, $P(g)$, $|(\mathscr{G} \circ \mathscr{D})^{-1}(g)|$ and $|\mathscr{F}'^{-1}(f)|$ based on Eqs. (3.30), (3.32), (3.18), (3.20) and (3.34), respectively.

The d-, g- and f-sequences when n is 2, 3, and 4 are shown along with the corresponding ranked tree shape and the four shape statistics, namely, $\beth = \beth(f)$, $\urcorner = \urcorner(f)$, $\widehat{\beth} = \widehat{\beth}(f)$ and $\ddot{f} = \ddot{\Lambda}(f)$ in Table 1.

When $n = 3$ we tabulate the state spaces, (backward) transition diagrams, the sequences and their probabilities at six resolutions of the n-coalescent in Table 4.

Example 1 (3 Samples) When there are 3 samples, we have 3 b-sequences, 3 c-sequences, 1 d-sequence, 1 g-sequence and 1 f-sequence. In Table 4, we tabulate the state-space, (backward) transition diagram, sequences and the corresponding probabilities at each of the six n-coalescent resolutions.

There is only one f-sequence, with $\urcorner(f) = 1$, $\ddot{\Lambda}(f) = \ddot{f} = (1,1)$ and $\prod_{i=2}^{3} \ddot{f}_i = 1$. Thus, $P(f) = (2^1/(3-1)!) = 1$ and $|\mathscr{F}^{-1}(f)| = 3!\ 2^{1+1-3} = 3$. Again, there is

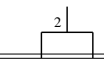
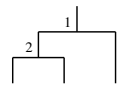
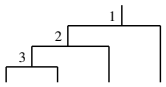
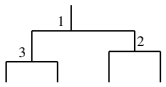
n	ranked tree shape	d-sequence	g-sequence	f-sequence	$\beth$	$\daleth$	$\widehat{\beth}$	$\ddot{f}$
2		$d=\begin{pmatrix}2\\0\end{pmatrix}$	$g=\begin{pmatrix}1\\0\end{pmatrix}$	$f=\begin{pmatrix}0&1\\2&0\end{pmatrix}$	1	0	0	(1)
3		$d=\begin{pmatrix}3&0\\0&2\\0&0\end{pmatrix}$	$g=\begin{pmatrix}1&0\\0&1\\0&0\end{pmatrix}$	$f=\begin{pmatrix}0&0&1\\1&1&0\\3&0&0\end{pmatrix}$	1	1	0	(1,1)
4		$d^{\curlywedge}=\begin{pmatrix}4&0&0\\0&3&0\\0&0&2\\0&0&0\end{pmatrix}$	$g^{\curlywedge}=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\\0&0&0\end{pmatrix}$	$f^{\curlywedge}=\begin{pmatrix}0&0&0&1\\1&0&1&0\\2&1&0&0\\4&0&0&0\end{pmatrix}$	1	2	0	(1,1,1)
4		$d^{\curlyvee}=\begin{pmatrix}4&0&0\\0&2&2\\0&0&2\\0&0&0\end{pmatrix}$	$g^{\curlyvee}=\begin{pmatrix}1&0&0\\0&1&1\\0&0&1\\0&0&0\end{pmatrix}$	$f^{\curlyvee}=\begin{pmatrix}0&0&0&1\\0&2&0&0\\2&1&0&0\\4&0&0&0\end{pmatrix}$	2	0	1	(1,2,1)

Table 1 The ranked tree shape and shape statistics of d-, g- and f-sequences when n is 2, 3, and 4.

only one f- and g-sequence with one cherry, i.e. $\beth(f)=\beth(g)=1$, and $\widehat{\beth}(f):=n-1-\daleth(f)-\beth(f)=3-1-1-1=0$. Therefore, $P(g)=2^{3-1-1}/(3-1)!=1$, $|(\mathscr{G}\circ\mathscr{D})^{-1}(g)|=3!\,2^{-1}=3$ and $|\mathscr{F}'^{-1}(f)|=2^{-0}\,1=1$.

Example 2 (4 Samples) In the case of four samples, there are 18 b-sequences, 18 c-sequences, 2 d-sequence, 2 g-sequence and 2 f-sequence. We provide the d-, g- and f-sequences in Table 1. Out of the 18 c-sequences in $\mathscr{C}_4$, it is possible to apply Eq. (3.23) and find that 12 c-sequences map to $f^{\curlywedge}$ and 6 map to $f^{\curlyvee}$. Note that the ranked tree shapes corresponding to all the c-sequences $\mathscr{F}^{-1}(f^{\curlywedge})$ is the completely unbalanced g-sequence $g^{\curlywedge}$ and that corresponding to all the c-sequences $\mathscr{F}^{-1}(f^{\curlyvee})$ is the completely balanced g-sequence $g^{\curlyvee}$. Finally, the shape statistic triple for the two f-sequences are:

$$(\beth(f^{\curlywedge}),\daleth(f^{\curlywedge}),\widehat{\beth}(f^{\curlywedge}))=(1,2,0)\quad\text{and}\quad(\beth(f^{\curlyvee}),\daleth(f^{\curlyvee}),\widehat{\beth}(f^{\curlyvee}))=(2,0,1)\ .$$

Let us examine the two f-sequences closely. For $f^{\curlyvee}$ with $\daleth(f^{\curlyvee})=0$, $\ddot{\Lambda}(f^{\curlyvee})=\ddot{f}^{\curlyvee}=(1,2,1)$ and $\prod_{i=2}^{4}\ddot{f}_i^{\curlyvee}=2$ we obtain $P(f^{\curlyvee})=(2^0/(4-1)!).2=1/3$, $|\mathscr{F}^{-1}(f^{\curlyvee})|=4!\,2^{0+1-4}\times 2=6$ and $|\mathscr{F}'^{-1}(f^{\curlyvee})|=2^{-1}\times 2=1$. Similarly, for $f^{\curlywedge}$ with $\daleth(f^{\curlywedge})=2$, $\ddot{\Lambda}(f^{\curlywedge})=\ddot{f}^{\curlywedge}=(1,1,1)$ and $\prod_{i=2}^{4}\ddot{f}_i^{\curlywedge}=1$, we obtain $P(f^{\curlywedge})=(2^2/(4-1)!)=2/3$, $|\mathscr{F}^{-1}(f^{\curlywedge})|=4!\,2^{2+1-4}=12$ and $|\mathscr{F}'^{-1}(f^{\curlywedge})|=2^{-0}=1$.

Let us examine the two g-sequences closely. For $g^{\curlyvee}$ with $\beth(g^{\curlyvee})=2$, $P(g^{\curlyvee})=2^{4-1-2}/(4-1)!=1/3$ and $|(\mathscr{G}\circ\mathscr{D})^{-1}(g^{\curlyvee})|=4!\,2^{-2}=6$ and for $g^{\curlywedge}$ with $\beth(g^{\curlywedge})=1$, $P(g^{\curlywedge})=2^{4-1-1}/(4-1)!=2/3$ and $|(\mathscr{G}\circ\mathscr{D})^{-1}(g^{\curlywedge})|=4!\,2^{-1}=12$.

Example 3 (5 Samples) In the case of five samples, there are 180 b-sequences, 180 c-sequences, 5 d-sequence, 5 g-sequence and 4 f-sequence. As shown in Table 2, we denote the g-sequences as g^{a}, g^{b}, g^{c}, g^{d} and g^{e}, the d-sequences as d^{a}, d^{b}, d^{c}, d^{d} and d^{e}, and the f-sequences as f^{a}, f^{b}, f^{cd} and f^{e}. Note that g^{c} and g^{d} as well as d^{c}

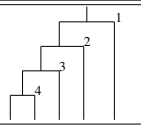
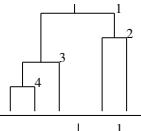
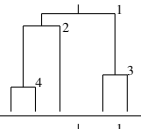
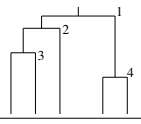
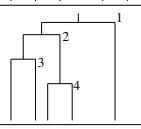
ranked tree shape	d-sequence	g-sequence	f-sequence	ℶ	ℸ	ℷ	$\ddot{f}$
	$d^{\mathtt{a}}=\begin{pmatrix}5&0&0&0\\0&4&0&0\\0&0&3&0\\0&0&0&2\\0&0&0&0\end{pmatrix}$	$g^{\mathtt{a}}=\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{pmatrix}$	$f^{\mathtt{a}}=\begin{pmatrix}0&0&0&0&1\\1&0&0&1&0\\2&0&1&0&0\\3&1&0&0&0\\5&0&0&0&0\end{pmatrix}$	1	3	0	$(1,1,1,1)$
	$d^{\mathtt{b}}=\begin{pmatrix}5&0&0&0\\0&2&3&0\\0&0&3&0\\0&0&0&2\\0&0&0&0\end{pmatrix}$	$g^{\mathtt{b}}=\begin{pmatrix}1&0&0&0\\0&1&1&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{pmatrix}$	$f^{\mathtt{b}}=\begin{pmatrix}0&0&0&0&1\\0&1&1&0&0\\2&0&1&0&0\\3&1&0&0&0\\5&0&0&0&0\end{pmatrix}$	2	2	0	$(1,1,1,1)$
	$d^{\mathtt{c}}=\begin{pmatrix}5&0&0&0\\0&3&2&0\\0&0&2&2\\0&0&0&2\\0&0&0&0\end{pmatrix}$	$g^{\mathtt{c}}=\begin{pmatrix}1&0&0&0\\0&1&1&0\\0&0&1&1\\0&0&0&1\\0&0&0&0\end{pmatrix}$	$f^{\mathtt{cd}}=\begin{pmatrix}0&0&0&0&1\\0&1&1&0&0\\1&2&0&0&0\\3&1&0&0&0\\5&0&0&0&0\end{pmatrix}$	2	2	0	$(1,1,2,1)$
	$d^{\mathtt{d}}=\begin{pmatrix}5&0&0&0\\0&3&0&2\\0&0&2&2\\0&0&0&2\\0&0&0&0\end{pmatrix}$	$g^{\mathtt{d}}=\begin{pmatrix}1&0&0&0\\0&1&0&1\\0&0&1&1\\0&0&0&1\\0&0&0&0\end{pmatrix}$	$f^{\mathtt{cd}}=\begin{pmatrix}0&0&0&0&1\\0&1&1&0&0\\1&2&0&0&0\\3&1&0&0&0\\5&0&0&0&0\end{pmatrix}$	2	2	0	$(1,1,2,1)$
	$d^{\mathtt{e}}=\begin{pmatrix}5&0&0&0\\0&4&0&0\\0&0&2&2\\0&0&0&2\\0&0&0&0\end{pmatrix}$	$g^{\mathtt{e}}=\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&1\\0&0&0&1\\0&0&0&0\end{pmatrix}$	$f^{\mathtt{e}}=\begin{pmatrix}0&0&0&0&1\\1&0&0&1&0\\1&2&0&0&0\\3&1&0&0&0\\5&0&0&0&0\end{pmatrix}$	2	1	1	$(1,1,2,1)$

Table 2 The d-, g- and f-sequences when $n=5$ are shown along with the corresponding ranked tree shape and the four shape statistics, namely, $ℶ=ℶ(f)$, $ℸ=ℸ(f)$, $\widehat{ℷ}=\widehat{ℷ}(f)$ and $\ddot{f}=\ddot{\Lambda}(f)$. Note that the third and forth row have the same f-sequence.

and $d^{\mathtt{d}}$ map to the same f-sequence $f^{\mathtt{cd}}$. Finally, the shape statistic triples for the four f-sequences are:

$$(ℶ(f^{\mathtt{a}}),ℸ(f^{\mathtt{a}}),\widehat{ℷ}(f^{\mathtt{a}}))=(1,3,0),\qquad (ℶ(f^{\mathtt{b}}),ℸ(f^{\mathtt{b}}),\widehat{ℷ}(f^{\mathtt{b}}))=(2,1,1),$$
$$(ℶ(f^{\mathtt{cd}}),ℸ(f^{\mathtt{cd}}),\widehat{ℷ}(f^{\mathtt{cd}}))=(2,2,0),\qquad (ℶ(f^{\mathtt{e}}),ℸ(f^{\mathtt{e}}),\widehat{ℷ}(f^{\mathtt{e}}))=(2,2,0).$$

For the four f-sequences: $f^{\mathtt{a}}$, $f^{\mathtt{b}}$, $f^{\mathtt{cd}}$ and $f^{\mathtt{e}}$, and the five g-sequences: $g^{\mathtt{a}}$, $g^{\mathtt{b}}$, $g^{\mathtt{c}}$, $g^{\mathtt{d}}$ and $g^{\mathtt{e}}$, we apply their shape statistics:

$$ℸ(f^{\mathtt{a}})=3\qquad ℸ(f^{\mathtt{b}})=1\qquad ℸ(f^{\mathtt{cd}})=ℸ(f^{\mathtt{e}})=2$$
$$ℶ(g^{\mathtt{a}})=1\qquad ℶ(g^{\mathtt{b}})=ℶ(g^{\mathtt{c}})=ℶ(g^{\mathtt{d}})=ℶ(g^{\mathtt{e}})=2$$
$$\prod_{i=2}^{5}\ddot{f}_i^{\mathtt{a}}=\prod_{i=2}^{5}\ddot{f}_i^{\mathtt{e}}=1^4=1\qquad \prod_{i=2}^{5}\ddot{f}_i^{\mathtt{b}}=\prod_{i=2}^{5}\ddot{f}_i^{\mathtt{cd}}=2,$$

to obtain the probabilities and cardinalities as follows:

$$\begin{aligned}
P(f^{\mathsf{a}}) = (2^3/(5-1)!) = P(f^{\mathsf{cd}}) = (2^2/(5-1)!) \times 2 &= 1/3 \\
P(f^{\mathsf{b}}) = (2^1/(5-1)!) \times 2 = P(f^{\mathsf{e}}) = (2^2/(5-1)!) &= 1/6 \\
|\mathscr{F}^{-1}(f^{\mathsf{a}})| = 5!\, 2^{3+1-5} = |\mathscr{F}^{-1}(f^{\mathsf{cd}})| = 5!\, 2^{2+1-5} \times 2 &= 60 \\
|\mathscr{F}^{-1}(f^{\mathsf{b}})| = 5!\, 2^{1+1-5} \times 2 = |\mathscr{F}^{-1}(f^{\mathsf{e}})| = 5!\, 2^{2+1-5} &= 30 \\
P(g^{\mathsf{a}}) = 2^{5-1-1}/(5-1)! &= 1/3 \\
P(g^{\mathsf{b}}) = P(g^{\mathsf{c}}) = P(g^{\mathsf{d}}) = P(g^{\mathsf{e}}) = 2^{5-1-2}/(5-1)! &= 1/6 \\
|(\mathscr{G} \circ \mathscr{D})^{-1}(g^{\mathsf{a}})| = 5!\, 2^{-1} &= 60 \\
|(\mathscr{G} \circ \mathscr{D})^{-1}(g^{\mathsf{b}})| = |(\mathscr{G} \circ \mathscr{D})^{-1}(g^{\mathsf{c}})| = 5!2^{-2} &= 30 \\
|(\mathscr{G} \circ \mathscr{D})^{-1}(g^{\mathsf{d}})| = |(\mathscr{G} \circ \mathscr{D})^{-1}(g^{\mathsf{e}})| = 5!\, 2^{-2} &= 30 \\
|\mathscr{F}'^{-1}(f^{\mathsf{a}})| = |\mathscr{F}'^{-1}(f^{\mathsf{e}})| = 2^{-0} &= 1 \\
|\mathscr{F}'^{-1}(f^{\mathsf{b}})| = 2^{-1} \times 2 &= 1 \\
|\mathscr{F}'^{-1}(f^{\mathsf{cd}})| = 2^{-0} \times 2 &= 2.
\end{aligned}$$

Applications of Eqs. (3.18) and (3.20) to the g-sequences of Examples 1, 2 and 3 above are consistent with those of Tajima's evolutionary relationships (Tajima, 1983, Figs. 1-3).

4 Why lump?

4.1 Nature and extent of Markov lumpings

Here we give some arguments showing the efficiency of looking at the appropriate resolution as discussed in Sect. 1.2. As a start, let us consider statistics of ranked labeled trees. We have seen that there is a bijection from $\mathscr{B}_n$, the set of b-sequences, as well as from $\mathscr{C}_n$, the set of c-sequences, to the set of ranked labeled trees. We introduced b-sequences since there are Markov lumpings from b-sequences to all other resolutions, but since the space $\mathbb{C}_n$ is much smaller than $\mathbb{B}_n$ (there are no vintage tags), we will only consider c-sequences when the object of interest in inference is a ranked labeled tree.

Let us first gain some insight on the extent of lumpings between $\mathbb{C}_n$, $\mathbb{G}_n$ and $\mathbb{F}_n$. Note that the cardinality of $\mathbb{C}_n$, $|\mathbb{C}_n|$, is the n-th Bell number in Eq. (3.9). Further, the cardinality of $\mathbb{G}_n$, $|\mathbb{G}_n|$, is the $(n+1)$-th Fibonacci number in Eq. (3.19). The cardinality of $\mathbb{F}_n$, $|\mathbb{F}_n|$, is the number of integer partitions of n in Eq. (3.22). The approximate values of $|\mathbb{C}_n|$, $|\mathbb{G}_n|$ and $|\mathbb{F}_n|$ are shown in Table 3 for typical samples sizes of interest to us. In fact, $|\mathbb{F}_n|/|\mathbb{G}_n| \to 0$ and $|\mathbb{G}_n|/|\mathbb{C}_n| \to 0$ as $n \to \infty$. Furthermore, Props. 10 and 12 precisely describe the number of c-sequences or b-sequences or ranked labeled trees that are coarsened into any specific g- or f-sequence, respectively. This can be advantageous during integrations, involving dynamic programming or sequential particle filtering algorithms (Del Moral, 2004; Doucet and Johansen, 2009), over paths of the Markov chain on $\mathbb{G}_n$ or $\mathbb{F}_n$ instead of $\mathbb{C}_n$ or over paths on $\mathbb{F}_n$ instead of $\mathbb{G}_n$, provided the coarser resolution preserves the likelihood of the statistic of interest. That

is, the sampling distribution of this statistic should depend on the (hidden) c-sequence only through its coarsening $\mathscr{F}(c) = f$ or $\mathscr{G}(c) = g$. Coming back to Eq. (1.1), we can then write for instance (extending the notation of Sect. 1.2 in an obvious way)

$$P(x_{obs}|\phi) = \int_{\mathscr{F}_n \otimes \mathbb{R}_+^n} P(x_{obs}|f \otimes t, \phi) p_\phi(f \otimes t)\, d\mathbb{P}(f \otimes t),$$

where this time the integration is carried out over the much smaller space $\mathscr{F}_n \otimes \mathbb{R}_+^n$ instead of $\mathscr{C}_n \otimes \mathbb{R}_+^n$.

Table 3 Cardinalities of the state spaces $\mathbb{C}_n$, $\mathbb{G}_n$ and $\mathbb{F}_n$.

$n = \lvert\mathbb{H}_n\rvert$	4	10	30	60	90	120
$\lvert\mathbb{C}_n\rvert$	15	1.2×10^5	8.5×10^{23}	9.8×10^{59}	1.4×10^{101}	5.1×10^{145}
$\lvert\mathbb{G}_n\rvert$	5	88	1.3×10^6	2.5×10^{12}	4.7×10^{18}	8.7×10^{24}
$\lvert\mathbb{F}_n\rvert$	5	42	5.6×10^3	9.7×10^5	5.7×10^7	1.8×10^9

Let us now turn to statistics of ranked tree shapes. As we have seen earlier, there is a bijection from $\mathscr{D}_n$, the set of d-sequences, as well as from $\mathscr{G}_n$, the set of g-sequences, to the set of ranked tree shapes. Again, since the state space of g-sequences is much smaller (as we do not track the size of components), we will only consider g-sequences when the object of interest in inference is a ranked tree shape. Moreover, for various shape statistics (of ranked tree shapes) whose likelihood only depends on the hidden f-sequence (as described below in Sect. 4.2), it is preferable to study the lumped Markov chain on $\mathbb{F}_n$ as opposed to that on $\mathbb{G}_n$ (since here again f-sequences contain the minimal information required to reconstruct the statistic of interest). Recall that Prop. 13 gives the number of g-sequences or d-sequences or ranked tree shapes that are coarsened into any specific f-sequence.

We now provide a few examples of statistics for which the appropriate coarsening of Kingman's coalescent may be used to simplify the desired computations.

4.2 Shape statistics where f-sequences are sufficient

We show that any f-sequence f realized under the unvintaged and sized n-coalescent captures a considerable amount of information about the ranked tree shapes in the equivalence class of c-sequences $\mathscr{F}^{-1}(f)$ or in $\mathscr{G}(\mathscr{F}'^{-1}(f))$. For instance, various tree shape statistics are further summaries of the f-sequence. We will make the former sentence precise by showing that several tree-shape statistics in the literature are functions of a sequence of $n-1$ ordered pairs obtained from f-sequences.

For a given c-sequence $c := (c_n, c_{n-1}, \ldots, c_1)$, the corresponding *Aldous' shape statistic sequence* (Aldous, 2001) or *$\tilde{s}$-sequence* is defined as $\tilde{s} := (\tilde{s}_n, \tilde{s}_{n-1}, \ldots, \tilde{s}_1)$, where $\tilde{s}_i := (\tilde{s}_{i,1}, \tilde{s}_{i,2})$. The i-th ordered pair $(\tilde{s}_{i,1}, \tilde{s}_{i,2})$ of the $\tilde{s}$-sequence is the size of the block $c_{i-1,j}$ that was created at the end of the i-th coalescent epoch and the size of the smaller of the two blocks that coalesced to create $c_{i-1,j}$. The $\tilde{s}$-sequence

is then obtained directly from an f-sequence f through the following mapping $\widetilde{S}$: $(f_n, f_{n-1}, \ldots, f_1) \mapsto \tilde{s} := (\tilde{s}_n, \tilde{s}_{n-1}, \ldots, \tilde{s}_2)$, with

$$\begin{aligned} \tilde{s}_i &:= (\tilde{s}_{i,1}, \tilde{s}_{i,2}) \\ &:= \left(\max(Df_i), \max(Df_i) + \min(Df_i) 2^{-\mathbb{1}_{\{0\}}(\max(Df_i)+2\min(Df_i))}\right), \\ \text{where} \quad Df_i &:= \{\, j(f_{i-1,j} - f_{i,j}) \;:\; j \in \{1,2,\ldots,n\} \,\}. \end{aligned} \tag{4.1}$$

In words, Df_i is the i-th "difference" vector whose j-th coordinate is 0 if no block of size j is created or coalesces between the i-th and $(i-1)$-st coalescent epochs, equals j if a block of size j is created during this step, equals $-j$ if a block of size j coalesces with a block of different size during this step, or equals $-2j$ if two blocks of size j coalesced between the i-th and $(i-1)$-st epochs (hence the factor $2^{-\mathbb{1}_{\{0\}}(\max(Df_i)+2\min(Df_i))}$ to obtain the size of the smallest block in this case). Therefore, f-sequences contain all the information on $\tilde{s}$-sequences. Aldous (2001) constructs the $\tilde{s}$-sequence forward in time using a tree-splitting model. This is partly motivated by a description of tree-shape imbalance via median-regression over a scatterplot of the ordered pairs $(\tilde{s}_{i,1}, \tilde{s}_{i,2})$'s obtained from phylogenetic trees that were estimated from DNA sequences of extant taxa (Aldous, 2001).

Next, let us show that several classical scalar-valued tree shape statistics are functions of $\tilde{s} = \widetilde{S}(f)$. First, consider the following family of statistics indexed by the nonempty subsets of $\{2,3,\ldots,n\}$:

$$\mathfrak{Q}_n := \left\{ Q_{\mathtt{I}} : \tilde{s} \mapsto \sum_{i=2}^{n} \tilde{s}_{i,1} \mathbb{1}_{\mathtt{I}}(\tilde{s}_{i,1}) \;:\; \mathtt{I} \subset \{2,3,\ldots,n\}, \mathtt{I} \neq \emptyset \right\}.$$

When $\mathtt{I} = \{2,\ldots,n\}$, $Q_{\{2,3,\ldots,n\}}(\tilde{s}) = \sum_{i=2}^{n} \tilde{s}_{i,1}$ is *Sackin's index* which is the sum of the number of leaves subtended by each internal node (Sackin, 1975; Tajima, 1983). Then, $Q_{\{2\}}/2$ is the *number of cherries*, i.e., the number of internal nodes that subtend exactly 2 leaves (McKenzie and Steel, 2000). There are $2^{n-1} - 3$ other scalar-valued shape statistics in the family $\mathfrak{Q}_n$ for the n-coalescent. Another scalar-valued statistic that needs more information than the number of leaves subtended by the set of internal nodes is the Colless' index (Colless, 1982). It is the sum of the absolute difference between the number of leaves subtended by the two branches bifurcating from each internal node up to a constant factor. The Colless' index of an f-sequence f only depends on its Aldous' shape statistic sequence $\widetilde{S}(f) = \tilde{s}$ and is given by $(n^2 - 3n + 2)^{-1} \sum_{i=2}^{n} (\tilde{s}_{i,1} - 2\tilde{s}_{i,2})$.

Thus, we have shown that any f-sequence f captures a lot of information about the ranked tree shapes in $\mathscr{G}(\mathscr{F}'^{-1}(f))$. However, some information is lost about the ranked tree shapes in the coarsening as one f-sequence may encode several distinct g-sequences (recall that 2 distinct g-sequences mapped to the same f-sequence in Example 3).

4.3 Shape Statistics where g-sequences are sufficient

In the previous section, we showed that sampling distributions of f-sequences are sufficient to obtain that of several tree shape statistics. However, there are statistics

based on ranked tree shapes for which the n-coalescent resolution of f-sequences is not sufficient. In Ford et al (2009), the *runs statistic* was proposed for detecting lineage-specific bursts within a population or between species.

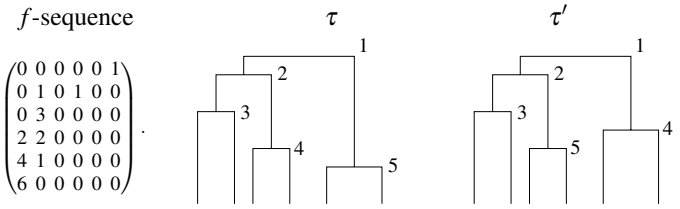


Fig. 3 Two ranked tree shapes on six leaves. Note that τ, the ranked tree shape in the middle panel, has runs statistic 4 while τ' on the right has runs statistic 5. However, both ranked tree shapes have the same f-sequence on the left.

The runs statistic is calculated recursively from a ranked tree shape τ. Note that the ranking on a tree shape is simply a total order of the interior vertices of the tree shape. By deleting the root of τ, we obtain two ranked tree shapes τ_1 and τ_2. The ranked tree shape τ is induced by these two ranked tree shapes τ_1 and τ_2 together with a *shuffle* on the interior vertices of τ_1 and τ_2. A shuffle puts the n_1 interior vertices in τ_1 and the n_2 interior vertices in τ_2 in order, e.g. 1112 means that first we have three bifurcations in τ_1, followed by one bifurcation in τ_2. The number of runs of a shuffle is the number of times we switch from i to j ($i \neq j$) plus one. Hence, our shuffle 1112 has two runs. The number of runs of a ranked tree shape τ is defined recursively by

$$R(\tau) = R(\tau_1) + R(\tau_2) + s(\tau) \ ,$$

where $s(\tau)$ is the number of runs in the shuffle on the interior vertices of τ_1 and τ_2. For further details, see Ford et al (2009).

As g-sequences can be mapped to ranked tree shapes via a bijection (Prop. 9), the g-sequences are sufficient for determining the runs statistic. On the other hand, the runs statistic cannot be obtained from f-sequences. For example, let us consider τ_1 and τ_2, the two ranked tree shapes in Fig. 3. There are 4 runs in τ (since $R(\tau) = R(\tau_1) + R(\tau_2) + s(\tau) = 2 + 0 + 2 = 4$) whereas τ' has 5 runs (since $R(\tau') = R(\tau'_1) + R(\tau'_2) + s(\tau') = 2 + 0 + 3 = 5$). However, both τ and τ' have the same f-sequence.

4.4 Statistics of Observed Mutations

Recall from Sect. 1.2 that a continuous coalescent tree $c \otimes t$, realized under the n-coalescent, describes the labeled ancestral history of the sampled individuals as a binary tree. Figure 4 shows a coalescent tree for a sample of four individuals. In the rest of this section, let us consider the following neutral models indexed by a parameter $\phi = (\phi_1, \phi_2) \in \boldsymbol{\Phi}$, in which mutations are independently super-imposed upon the coalescent trees at each site according to the infinitely-many-sites (IMS) model (Watterson, 1975). Under this model, independent mutations are super-imposed on

the coalescent tree $c \otimes t$ at each site according to a homogeneous Poisson process with rate ϕ_1, where $\phi_1 := 4N_e\mu$, N_e is the *effective population size*, and μ is the mutation rate per generation per site. We further stipulate that at most one mutation is allowed per site. The ancestral state is coded as 0 and the derived or mutant state is coded as 1.

The parameter ϕ_2 is the exponential growth rate of the population. That is, we assume that the whole population alive t units of time in the past was $e^{-\phi_2(t-s)}$ as big as the population alive $s < t$ units of time in the past. This parameter will appear in the computation of the likelihood of the continuous tree in Sect. 4.4.3. We keep the model simple with just two parameters in order to emphasize the advantages of working with the appropriate Markov lumping of the hidden genealogy space (depending on the statistics of the observed mutations that are being used to infer the parameters). But our ideas generalize in a straightforward manner to more complex demographic scenarios involving multiple independent loci that are evolving under the infinitely-many-sites model of mutation.

4.4.1 Binary Incidence Matrix.

Let us first describe the most precise observation of mutations we can obtain from the DNA sequence of n individuals at the present time under such a model. To this end, let us assume that the ancestral nucleotides are known, and that at most one derived nucleotide occurs at each site among the n sampled sequences (such bi-allelic data is common and sites showing both ancestral and derived characters are commonly referred to as *single nucleotide polymorphisms* or SNPs). Then from the aligned sequence data u, we obtain a *binary incidence matrix*, or *BIM*, $v \in \mathscr{V}_n^m := \{0,1\}^{n\times m}$ by replacing all ancestral states with 0 and derived states with 1.

BIM data is modeled by super-imposing the infinitely-many-sites model of mutation onto an n-coalescent sample genealogy (Kingman, 1982a,b). We can conduct inference on the basis of the observed BIM v using established importance sampling methods (Bahlo and Griffiths, 1996; Birkner and Blath, 2008; Griffiths and Tavare, 1994, 1996; Iorio and Griffiths, 2004; Slatkin, 2002; Stephens and Donnelly, 2000). However, the only resolutions of the coalescent containing all the information in BIM are the vintaged and unvintaged labeled n-coalescent. Inference based on the likelihood of BIM does not computationally scale well with realistic increases in the sample size and/or model complexity. Thus, computational population geneticists increasingly rely on summary statistics of BIM in order to conduct computationally feasible inference using simulation-intensive and "likelihood-free" methods such as ABC but with admittedly less information than that in the available BIM data. Hence, in this study we are not interested in direct inference on the basis of the observed BIM, but instead on further summary statistics of BIM.

4.4.2 Site Frequency Spectrum.

We can obtain the *site frequency spectrum* x from the BIM v via its *site sum spectrum* or *SSS* w. With w denoting the vector of column sums of v, the SFS x is the vector of frequencies of occurrences of each positive integer in w. Thus the i-th entry of x

records at how many sites exactly i sequences in u show the derived state (in other words, how many mutations are carried by exactly i individuals in our sample). We assume that no site displays only the derived state. Thus, x has only $n-1$ entries. Figure 4 depicts the BIM v, SSS w and SFS x on the right for a sample of four individuals with the genealogical and mutational history on the left.

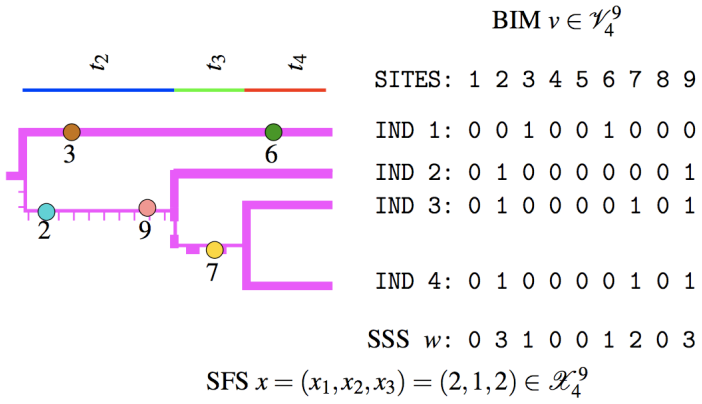


Fig. 4 At most one mutation per site under the infinitely-many-sites model are super-imposed as a homogeneous Poisson process upon the realization of identical coalescent trees at nine homologous sites labeled $\{1,2,\ldots,9\}$ that constitute a non-recombining locus from four individuals labeled $\{1,2,3,4\}$.

4.4.3 Inference based on statistics of observed mutations.

Let us now describe the basic probability models required to compute the likelihood of the SFS.

Recall from the beginning of Sect. 4.4 that the continuous parameter indexing our illustrative model is two-dimensional, i.e., the state-space $\boldsymbol{\Phi}$ satisfies $\boldsymbol{\Phi} := (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) \subset \mathbb{R}_+^2$. The second parameter ϕ_2 is the growth rate of our population, whose size is growing exponentially from the past. In our models, this translates into the property that at time t in the past, the instantaneous rate of coalescence of each pair of ancestral blocks is equal to $1/e^{-\phi_2 t}$, the inverse of the (relative) population size at this time. As a consequence, if there are i such blocks at that time, the instantaneous rate at which a pair of them coalesces is equal to $\binom{i}{2} e^{\phi_2 t}$. Calling T_i the length of the epoch during which there are i distinct ancestors to our sample (or blocks in the continuous-time coalescent), we thus have that for every $\tau_i, t_i \geq 0$,

$$\begin{aligned} P(T_i > t_i \,|\, T_n + \cdots + T_{i+1} = \tau_i, \phi_2) &= \exp\left\{ -\binom{i}{2} \int_{\tau_i}^{\tau_i + t_i} e^{\phi_2 u}\, du \right\} \\ &= \exp\left\{ -\binom{i}{2} \frac{1}{\phi_2} e^{\phi_2 \tau_i} \left(e^{\phi_2 t_i} - 1\right) \right\}, \end{aligned}$$

so that the Lebesgue density of the vector $(T_n,\dots,T_2)$ is given by

$$p_{\phi_2}(t_n,\dots,t_2)=\prod_{i=2}^{n}\left[\binom{i}{2}e^{\phi_2(t_n+\cdots+t_{i+1}+t_i)}\exp\left\{-\binom{i}{2}\frac{1}{\phi_2}e^{\phi_2\tau_i}\left(e^{\phi_2 t_i}-1\right)\right\}\right]. \quad (4.2)$$

Note that if $\phi_2=0$ (i.e., the population does not grow), we recover the density

$$p_0(t_n,\dots,t_2)=\prod_{i=2}^{n}\left[\binom{i}{2}e^{-\binom{i}{2}t_i}\right]$$

of the epoch times in the standard Kingman coalescent.

The first parameter ϕ_1 is the per-locus mutation rate scaled by the effective population size and is often denoted by θ in the population genetics literature. Once the (continuous-time) genealogical tree is realized, we assume that mutations fall on this tree at rate ϕ_1. Hence, if a given portion of the tree has length ℓ, the number of mutations it carries is a Poisson random variable with parameter $\phi_1\ell$.

This way, we have defined our family of models indexed by $(\phi_1,\phi_2)\in\mathbb{R}_+^2$. For Bayesian decisions, we allow our parameter to be a random vector $\Phi:=(\Phi_1,\Phi_2)$ with a Lebesgue-dominated density $p(\phi)$ and realizations $\phi:=(\phi_1,\phi_2)$. This prior density $p(\phi)$ is taken to be a uniform density over a compact rectangle to allow simple interpretations from Bayesian, frequentist and information-theoretic schools of inference. We are interested in approximately sufficient statistics in the sense of Le Cam (1964) for the purpose of computational efficiency. Recall that, informally, a statistic Z of the full experiment $E=(C,M)$ (say, the coalescent with mutations) is called *sufficient* if for every c,m and z,

$$P\big((C,M)=(c,m)\,|\,Z=z,\phi\big)$$

is independent of ϕ. In our models, this happens in particular when the statistic of interest depends on the full c-sequence only through a coarser resolution, say the corresponding f-sequence, on top of which mutations are super-imposed. In this case, we obtain Bayes' sufficiency in the sense of Kolmogorov (1942), in terms of the following posterior identity:

$$P(\phi|c)=P(\phi|\mathscr{F}(c)=f).$$

To carry on with our example, take the statistic of interest to be the full SFS X, which depends only on the information contained in f-sequences and on ϕ. For every $\phi\in\boldsymbol{\Phi}$ and $x\in\mathbb{Z}_+^{n-1}$, we have

$$\begin{aligned}P(\phi\,|\,X=x)&=\frac{P(X=x,\phi)}{P(X=x)}=\frac{\sum_{f\sim x}P(X=x,F=f,\phi)}{\sum_{f\sim x}\int_{\boldsymbol{\Phi}}P(X=x,F=f,\phi)d\phi}\\&=\frac{\sum_{f\sim x}P(X=x\,|\,F=f,\phi)P(F=f)p(\phi)}{\sum_{f\sim x}\int_{\boldsymbol{\Phi}}P(X=x\,|\,F=f,\phi)P(F=f)p(\phi)d\phi},\end{aligned}$$

where F is the random f-sequence describing the discrete-time genealogy, and $f\sim x$ means that the SFS x is compatible with the f-sequence f (so that x may arise if mutations are super-imposed on f). Since we can compute all the terms appearing in

the above r.h.s., there remains only to sum over all possible f-sequences to compute $P(\phi \,|\, X = x)$. As the number of f-sequences is much smaller than that of c-sequences even for not so large n's, the gain in this summation is huge compared to the classical formula where the sum is over all possible c-sequences. In the present case, the technique is all the more interesting if one is able to set up a controlled algorithm yielding only f-sequences that are compatible with the observed SFS (as in Sainudiin et al (2011)).

To finish with a concrete (though rather trivial) example, let us be less ambitious and consider the total number of mutations seen in the sample. That is, our new statistic is $S = \sum_{i=1}^{n-1} X_i$. This time, since we do not care about the size of the blocks on which the mutations fall, this statistic depends only on the lineage death process $\{H^{\uparrow}(t)\}$ and on ϕ. Since there is only one possible h-sequence, namely $h_* = \{n, n-1, \ldots, 1\}$, the total length of the tree is equal to $\sum_{i=2}^{n} iT_i$ (whose law depends on ϕ_2 only) and conditionally on this length, the number of mutations on the tree is a Poisson random variable with parameter $\phi_1 \sum_{i=2}^{n} iT_i$. Thus, we obtain that for every $s \in \mathbb{Z}_+$,

$$P(\phi \,|\, S = s) = \frac{P(S = s \,|\, H = h_*, \phi) P(H = h_*) p(\phi)}{\int_{\boldsymbol{\Phi}} P(S = s \,|\, H = h_*, \phi) P(H = h_*) p(\phi) d\phi}. \tag{4.3}$$

Now, recalling Eq. (4.2) we have

$$\begin{aligned} P(S = s \,|\, H = h_*, \phi) &= P\left(\text{Poisson}\left(\phi_1 \sum_{i=2}^{n} iT_i \right) = s \,\middle|\, \phi_2 \right) \\ &= \int_{\mathbb{R}_+^{n-1}} P\left(\text{Poisson}\left(\phi_1 \sum_{i=2}^{n} it_i \right) = s \right) p_{\phi_2}(t_n, \ldots, t_2)\, dt_n \cdots dt_2 \\ &= \int_{\mathbb{R}_+^{n-1}} e^{-(\phi_1 \sum_{i=2}^{n} it_i)} \frac{(\phi_1 \sum_{i=2}^{n} it_i)^s}{s!} p_{\phi_2}(t_n, \ldots, t_2)\, dt_n \cdots dt_2, \end{aligned}$$

and there only remains to plug this equality in the r.h.s. of Eq. (4.3) to obtain an expression for the probability of $\phi = (\phi_1, \phi_2)$ under the observed number of mutations s.

A strategy similar to those outlined above for computing the likelihood of the site frequency spectrum and the number of mutations can be applied to obtain the likelihood of other summary statistics by integrating over the appropriate and minimal coalescent resolution of hidden genealogies. Moreover, the Markov lumping maps and their inverse images along with the probabilities at each coalescent resolution described here allow for the design of novel sequential Monte Carlo algorithms, using particle systems (Del Moral, 2004; Doucet and Johansen, 2009) in the hidden genealogy spaces, that can consistently and adaptively move through increasingly refining coalescent resolutions in synchrony with increasingly refining statistics of the observed mutations.

Acknowledgements We are grateful to Robert C. Griffiths for his insights, comments and guidance on this project, to John Rhodes and Mike Steel for their comments on an earlier version of this manuscript, and to Mike Steel for pointing out (Kemeny and Snell, 1960, Defn. 6.3.1). During the initial course of this study, R.S. was supported by a research fellowship from the Royal Commission for the Exhibition

of 1851 and T.S. was supported by a PhD scholarship of the German Science Foundation and a summer studentship of the Allan Wilson Centre. A.V. was supported by the ANR project MANEGE (ANR-09-BLAN-0215) and T.S. by the Swiss National Science foundation. R.S. and A.V. were supported in part by the chaire Modélisation Mathématique et Biodiversité of Veolia Environnement-École Polytechnique-Museum National d'Histoire Naturelle-Fondation X.

References

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

Bahlo M, Griffiths R (1996) Inference from gene trees in a subdivided population. Theoret Pop Biol 57:79–95

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

Beaumont M, Robert C, Marin JM, Cornuet J (2009) Adaptivity for ABC algorithms: the ABC-PMC scheme. Biometrika 96(4):983–990

Birkner M, Blath J (2008) Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. J Math Biol 57:435–465

Colless DH (1982) Review of phylogenetics: the theory and practice of phylogenetic systematics. Systematic Zoology 31:100–104

Del Moral P (2004) Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Springer, New York

Doucet A, Johansen AM (2009) A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later. In: Crisan D, Rozovsky B (eds) The Oxford Handbook of Nonlinear Filtering, Oxford University Press

Etheridge AM (2011) Some mathematical models from population genetics. Lecture Notes in Math. 2012. Springer

Fisher R (1930) The Genetical Theory of Natural Selection. Clarenson, Oxford

Ford D, Matsen E, Stadler T (2009) A method for investigating relative timing information on phylogenetic trees. Syst Biol 58(2):167–183

Griffiths R, Tavare S (1994) Ancestral inference in population genetics. Stat Sci 9:307–319

Griffiths R, Tavare S (1996) Markov chain inference methods in population genetics. Math Comput Modelling 23:141–158

Iorio M, Griffiths R (2004) Importance sampling on coalescent histories. I. Adv Appl Prob 36:417–433

Kemeny J, Snell J (1960) Finite Markov chains. D. van Nostrand Company, Inc., Princeton

Kendall DG (1975) Some problems in mathematical genealogy. In: Gani J (ed) Perspectives in Probability and Statistics, Academic Press, pp 325–345

Kingman JFC (1982a) The coalescent. Stochastic Proc Appl 13:235–248

Kingman JFC (1982b) On the genealogy of large populations. J Applied Probab 19:27–43

Kolmogorov A (1942) Sur l'estimation statistique des paramètres de la loi de gauss. Bull Acad Sci URSS Ser Math 6:3–32

Le Cam L (1964) Sufficiency and approximate sufficiency. Ann Math Stats 35:1419–1455
Leuenberger C, Wegmann D (2009) Bayesian computation and model selection without likelihoods. Genetics 184:243–252
Marin JM, Pudlo P, Robert CP, Ryder RJ (2012) Approximate bayesian computational methods. Statistics and Computing 22(6):1167–1180, DOI 10.1007/s11222-011-9288-2
Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci USA 100:15,324–15,328
McKenzie A, Steel M (2000) Distribution of cherries for two models of trees. Math Biosci 164:81–92
Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16:1791–1798
Sackin MJ (1975) "Good" and "bad" phenograms. Systematic Zoology 21:225–226
Sainudiin R, Thornton K, Harlow J, Booth J, Stillman M, Yoshida R, Griffiths R, McVean G, Donnelly P (2011) Experiments with the site frequency spectrum. Bulletin of Mathematical Biology 73(4):829–872
Semple C, Steel M (2003) Phylogenetics. Oxford University Press
Sisson S, Fan Y, Tanaka M (2007) Sequential Monte Carlo without likelihoods. Proc Natl Acad Sci USA 104:1760–1765
Slatkin M (2002) A vectorized method of importance sampling with applications to models of mutation and migration. Theor Pop Biol 62:339–348
Stephens M, Donnelly P (2000) Inference in molecular population genetics. J R Statist Soc B 62:605–655
Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460
Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595
Tavaré S (1983) Line-of-descent and genealogical processes, and their applications in population genetics models. Theor Pop Biol 26:119–164
Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276
Weiss G, von Haeseler A (1998) Inference of population history using a likelihood approach. Genetics 149:1539–1546
Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

5 Appendix

Proof 1 (of Prop. 1) We first prove Eq. (3.1). When there are i vintaged lineages in the i-th coalescent epoch, a coalescence event can reduce the number of lineages to $i-1$ by coalescing one of $\binom{i}{2}$ many pairs of vintaged lineages uniformly at random. Hence, the inverse $\binom{i}{2}^{-1}$ appears in the transition probabilities. The conditions that $b_{i-1} \prec_b b_i$ and $b_i \in \mathbb{B}_n^i$ for each $i \in \{n, n-1, \ldots, 3, 2\}$ ensure that our b-sequence

$b = (b_n, \ldots, b_1)$ remains in $\mathscr{B}_n$ as we go backwards in time from the i-th coalescent epoch with i samples to the $(i-1)$-th coalescent epoch.

Next, we prove Eq. (3.4) and Eq. (3.2). Without loss of generality, let us chronologically list $b_i = \{c_{i,1}^{\langle m_{i,1}\rangle}, c_{i,2}^{\langle m_{i,2}\rangle}, \ldots, c_{i,i}^{\langle m_{i,i}\rangle}\}$, such that $m_{i,1} \leq m_{i,2} \leq \cdots \leq m_{i,i}$. Let $c_{i,1:j} := c_{i,1} \cup c_{i,2} \cup \cdots \cup c_{i,j}$, where the $c_{i,j}$'s are the unvintaged blocks in the chronologically listed b_i. For a b-sequence $b \in \mathbb{B}_n$ define $b_{n:i} := (b_n, b_{n-1}, \ldots, b_{i+1}, b_i)$. Then by Eq. (3.1) and the Markov property of $\{B^{\uparrow}(t)\}$,

$$P(b_{n:i}) := P((b_n, \ldots, b_i)) = P(b_{n-1}|b_n) \ldots P(b_i|b_{i+1}) = \frac{2^{n-i} i!(i-1)!}{n!(n-1)!} \quad .$$

In particular, each sequence $b_{n:i}$ is equally likely and the case $i = 1$ yields Eq. (3.4). Let N_i be the number of $b_{n:i}$-sequences which lead to $b_i = \{c_{i,1}^{\langle m_{i,1}\rangle}, c_{i,2}^{\langle m_{i,2}\rangle}, \ldots, c_{i,i}^{\langle m_{i,i}\rangle}\}$. Determining N_i establishes the probability for b_i, since each $b_{n:i}$-sequence is equally likely, i.e.

$$P(b_i) = P(b_{n:i}) N_i. \tag{5.1}$$

Recall that $i' = \max\{j : m_{i,j} < n\}$. For a vintaged lineage $b_{i,j} = c_{i,j}^{\langle m_{i,j}\rangle}$ in epoch i, the number of possible b-sequences in $\mathscr{B}_{|c_{i,j}|}$ using Eq. (3.4) is:

$$|\mathscr{B}_{|c_{i,j}|}| = \frac{|c_{i,j}|!(|c_{i,j}|-1)!}{2^{|c_{i,j}|-1}}. \tag{5.2}$$

In order to calculate N_i, let us define $N_{i,j}$ as the number of b-sequences on the label set $c_{i,1:j}$ stopped when all but j lineages coalesced, respecting (i) a fixed b-sequence on the label set $c_{i,1:j-1}$ stopped when all but $j-1$ lineages coalesced, and (ii) a fixed b-sequence on the label set $c_{i,j}$, $j \leq i'$ stopped when all lineages coalesced. We have

$$N_i = |\mathscr{B}_{|c_{i,1}|}| \times |\mathscr{B}_{|c_{i,2}|}| \times N_{i,2} \times |\mathscr{B}_{|c_{i,3}|}| \times N_{i,3} \times \cdots \times |\mathscr{B}_{|c_{i,i'}|}| \times N_{i,i'}. \tag{5.3}$$

We will now determine $N_{i,j}$. Note that there are $|c_{i,1:j-1}| - (j-1)$ coalescent events on $c_{i,1:j-1}$ up to epoch i. There are $|c_{i,j}| - 1$ coalescent events on $c_{i,j}$. The coalescent events in epoch $i, i+1, \ldots, m_{i,j} - 1$ happen on $c_{i,1:j-1}$, while by definition the coalescent event in epoch $m_{i,j}$ happens on $c_{i,j}$. The remaining elements are shuffled together arbitrarily, the number of possible shuffles equals $N_{i,j}$, which is,

$$N_{i,j} = \binom{|c_{i,1:j-1}| - (j-1) - (m_{i,j} - i) + |c_{i,j}| - 2}{|c_{i,j}| - 2}. \tag{5.4}$$

So overall, using Equations (5.1) – (5.4), we obtain,

$$
\begin{aligned}
P(b_i) &= P(b_{n:i})N_i \\
&= 2^{n-i}\,\frac{i!(i-1)!}{n!(n-1)!}\prod_{j=1}^{i'}\frac{|c_{i,j}|!(|c_{i,j}|-1)!}{2^{|c_{i,j}|-1}} \\
&\qquad \prod_{j=2}^{i'}\binom{|c_{i,1:j-1}|-(j-1)-(m_{i,j}-i)+|c_{i,j}|-2}{|c_{i,j}|-2} \\
&= \frac{i!(i-1)!}{n!(n-1)!}\left(\prod_{j=1}^{i'}|c_{i,j}|!\right)(|c_{i,1}|-1)!\left(\prod_{j=2}^{i'}\frac{(|c_{i,j}|-1)(|c_{i,1:j}|-j-1-(m_{i,j}-i))!}{(|c_{i,1:j-1}|-j+1-(m_{i,j}-i))!}\right) \\
&= \frac{i!(i-1)!}{n!(n-1)!}\left(\prod_{j=1}^{i'}|c_{i,j}|!\right)(|c_{i,1}|-1)! \\
&\qquad \left(\prod_{j=2}^{i'}(|c_{i,j}|-1)\right)\left(\frac{\prod_{j=2}^{i'}(|c_{i,1:j}|-j-1-(m_{i,j}-i))!}{\prod_{j=1}^{i'-1}(|c_{i,1:j}|-j-(m_{i,j+1}-i))!}\right) \\
&= \frac{i!(i-1)!}{n!(n-1)!}\left(\prod_{j=1}^{i'}|c_{i,j}|!(|c_{i,j}|-1)\right)\left(\frac{\prod_{j=1}^{i'}(|c_{i,1:j}|-j-1-m_{i,j}+i)!}{\prod_{j=1}^{i'-1}(|c_{i,1:j}|-j-m_{i,j+1}+i)!}\right) \\
&= \frac{i!(i-1)!}{n!(n-1)!}\left(\frac{\prod_{j=1}^{i'}|c_{i,j}|!(|c_{i,j}|-1)(|c_{i,1:j}|-j-1-m_{i,j}+i)!}{\prod_{j=1}^{i'-1}(|c_{i,1:j}|-j-m_{i,j+1}+i)!}\right)
\end{aligned}
$$

which completes the proof of Eq. (3.2).

Finally, Eq. (3.3) is obtained by Bayes' rule.

Proof 2 (of Eqs. (3.7) and (3.6)) First, Eq. (3.7) is a direct application of Bayes' rule to Eq. (3.5) and Eq. (3.6). Indeed, if $c_{i-1} \prec_c c_i$ with $c_i \in \mathbb{C}_n^i$ and j_*, j'_*, j''_* such that $c_{i,j_*} \cup c_{i,j'_*} = c_{i-1,j''_*} \in c_{i-1}$, then

$$
\begin{aligned}
P(c_i|c_{i-1}) &= \frac{P(c_{i-1}|c_i)P(c_i)}{P(c_{i-1})} \\
&= \frac{(n-i)!\,i!\,(i-1)!\prod_{j=1}^{i}|c_{i,j}|!\,n!\,(n-1)!}{\binom{i}{2}n!\,(n-1)!(n-i+1)!\,(i-1)!\,(i-2)!\prod_{j=1}^{i-1}|c_{i-1,j}|!} \\
&= \frac{2\prod_{j=1}^{i}|c_{i,j}|!}{(n-i+1)\prod_{j=1}^{i-1}|c_{i-1,j}|!} = \frac{2\,|c_{i,j_*}|!\,|c_{i,j'_*}|!}{(n-i+1)|c_{i-1,j''_*}|!} \\
&= \frac{2\,|c_{i,j_*}|!\,|c_{i,j'_*}|!}{(n-i+1)(|c_{i,j_*}|+|c_{i,j'_*}|)!} = \frac{2}{(n-i+1)\binom{|c_{i,j_*}|+|c_{i,j'_*}|}{|c_{i,j_*}|}}.
\end{aligned}
$$

If we do not have $c_{i-1} \prec_c c_i$, then $P(c_i|c_{i-1}) = 0$.

Next, for completeness and to exemplify the coarsening relating b- and c-sequences, we show that $P(c_i)$ can also be obtained from $P(b_i)$ in Eq. (3.2). Since we are not interested in the coalescent vintage of any of our lineages, the quantity of interest in Eq. (5.4) becomes

$$
\binom{|c_{i,1:j-1}|-(j-1)+|c_{i,j}|-1}{|c_{i,j}|-1}
$$

as we allow any shuffle of the $|c_{i,1:j-1}|-(j-1)$ coalescent events with the $|c_{i,j}|-1$ coalescent events. Let $i'=\max\{j: m_{i,j}<n\}$. We have,

$$\begin{aligned}
P(c_i) &= 2^{n-i}\frac{i!(i-1)!}{n!(n-1)!}\prod_{j=1}^{i'}\frac{|c_{i,j}|!(|c_{i,j}|-1)!}{2^{|c_{i,j}|-1}}\prod_{j=2}^{i'}\binom{|c_{i,1:j-1}|-(j-1)+|c_{i,j}|-1}{|c_{i,j}|-1}\\
&= \frac{i!(i-1)!}{n!(n-1)!}\left(\prod_{j=1}^{i'}|c_{i,j}|!\right)(|c_{i,1}|-1)!\left(\prod_{j=2}^{i'}\frac{(|c_{i,1:j}|-j)!}{(|c_{i,1:j-1}|-(j-1))!}\right)\\
&= \frac{i!(i-1)!}{n!(n-1)!}\left(\prod_{j=1}^{i'}|c_{i,j}|!\right)\left(\frac{\prod_{j=1}^{i'}(|c_{i,1:j}|-j)!}{\prod_{j=1}^{i'-1}(|c_{i,1:j}|-j)!}\right)\\
&= \frac{i!(i-1)!}{n!(n-1)!}\left(\prod_{j=1}^{i'}|c_{i,j}|!\right)(n-i)!
\end{aligned}$$

since $(|c_{i,1:i'}|-i')!=(n-(i-i')-i')!$ and $|c_{i,j}|=1$ if $m_{i,j}=n$. Therefore, $P(c_i)$ can be obtained from $P(b_i)$, the probability that a vintaged and labeled n-coalescent visits a particular vintaged partition b_i in $\mathbb{B}_n^i$.

Proof 3 (of Prop. 5) We first prove Eq. (3.10). The number of leaf lineages that coalesced at the end of epoch i is $d_{i,n}-d_{i',n}$, where $\mathbb{D}_n^{i-1}\ni d_{i'}\prec_d d_i\in\mathbb{D}_n^i$. Note that $(d_{i,n}-d_{i',n})\in\{0,1,2\}$, for any $i\in\{2,3,\ldots,n\}$. Therefore, three type of coalescent events need to be discriminated among the $\binom{i}{2}$ many pairs from i distinct lineages during epoch i. First, when $(d_{i,n}-d_{i',n})=0$ we have a coalescent event between two specific non-leaf lineages, each with coalescent vintage smaller than n. Thus, there is exactly $\binom{d_{i,n}}{0}=1$ such event among $\binom{i}{2}$ possibilities. Second, when $(d_{i,n}-d_{i',n})=1$ we have a coalescent event between one specific non-leaf lineage and any one of $d_{i,n}$ many leaf lineages (which we do not discriminate at this resolution). Thus, there are exactly $\binom{d_{i,n}}{1}=d_{i,n}$ many events among $\binom{i}{2}$ possibilities of the second type. Third, when $(d_{i,n}-d_{i',n})=2$ we have a coalescent event between any two of $d_{i,n}$ many leaf lineages. Thus, there are exactly $\binom{d_{i,n}}{2}$ many events among $\binom{i}{2}$ possibilities of the third type. All three types of events are accounted for in Eq. (3.10).

Next we prove that the probability of $\{D^{\uparrow}(k)\}_{k\in[n]_-}$ visiting a state $d_i\in\mathbb{D}_n^i$ is given by Eq. (3.11), where $d_{i,1:j}:=\sum_{k=1}^{j}d_{i,k}$ and $m'_{i,j}:=\min\{k>j: d_{i,k}>0\}$ and $k_{i,j}:=|\{m\le j: d_{i,m}>0\}|$. We exploit the Markov lumping from $\mathbb{B}_n$ to $\mathbb{D}_n$ and derive $P(d_i)$ from $P(b_i)$ in Eq. (3.2), where $d_i\in\mathbb{D}_n^i$ and $b_i\in\mathbb{B}_n^i$ such that dropping the labels in each subset of b_i (but retaining the size and vintage) yields $d_i=\mathscr{D}(b_i)$. We count the number of possible labelings of an element d_i. This is $\frac{n!}{\prod_{j=1}^{n-1}d_{i,j}!}$. Let b_i be such a labeling. By Eq. (3.2) and the fact that any partial b-sequence is equally likely to occur (see Proof 1), we have

$$\begin{aligned}
P(d_i) &= P(b_i)\frac{n!}{\prod_{j=1,d_{i,j}>0}^{n-1}d_{i,j}!}\\
&= \frac{i!(i-1)!}{n!(n-1)!}\left(\frac{\prod_{j=1,d_{i,j}>0}^{n-1}d_{i,j}!(d_{i,j}-1)(d_{i,1:j}-k_{i,j}-1-j+i)!}{\prod_{j=1,d_{i,j}>0}^{n-1}(d_{i,1:j}-k_{i,j}-m'_{i,j}+i)!}\right)\frac{n!}{\prod_{j=1,d_{i,j}>0}^{i}d_{i,j}!}\\
&= \frac{i!(i-1)!}{(n-1)!}\left(\frac{\prod_{j=1,d_{i,j}>0}^{n-1}(d_{i,j}-1)(d_{i,1:j}-k_{i,j}-j-1+i)!}{\prod_{j=1,d_{i,j}>0}^{n-1}(d_{i,1:j}-k_{i,j}-m'_{i,j}+i)!}\right).
\end{aligned}$$

From Eq. (3.10) and Eq. (3.11) we get Eq. (3.12) by Bayes rule.

Finally we prove Eq. (3.13).

$$
\begin{aligned}
P(d) &= \prod_{i=2}^{n} P(d_{i-1}|d_i) = \prod_{i=2}^{n} \binom{d_{i,n}}{d_{i,n}-d_{i-1,n}} \binom{i}{2}^{-1} \\
&= \prod_{i=2}^{n} \frac{d_{i,n}!}{d_{i-1,n}!(d_{i,n}-d_{i-1,n})!} \binom{i}{2}^{-1} = d_{n,n}! \prod_{i=2}^{n} \frac{1}{(d_{i,n}-d_{i-1,n})!} \binom{i}{2}^{-1} \\
&= n! \left(\prod_{i=2}^{n} ((d_{i,n}-d_{i-1,n})!)^{-1} \right) \left(\prod_{i=2}^{n} \binom{i}{2}^{-1} \right) \\
&= n! \left(\prod_{j=0,1,2} (j!)^{-\sum_{i=2}^{n} \mathbf{1}_{\{j\}}(d_{i,n}-d_{i-1,n})} \right) \prod_{i=2}^{n} \binom{i}{2}^{-1} \\
&= n! \left(1 \times 1 \times 2^{-\sum_{i=2}^{n} \mathbf{1}_{\{2\}}(d_{i,n}-d_{i-1,n})} \right) \prod_{i=2}^{n} \binom{i}{2}^{-1} \\
&= \frac{n!}{2^{\mathfrak{I}(d)}} \prod_{i=2}^{n} \binom{i}{2}^{-1} = \frac{2^{n-\mathfrak{I}(d)-1}}{(n-1)!} \quad .
\end{aligned}
$$

Proof 4 (of Prop. 7) First we prove that the transition probability of the jump Markov chain $\{G^{\uparrow}(k)\}_{k\in[n]_-}$ on $\mathbb{G}_n$ is given by Eq. (3.15), where $g_{i,n}$ is the number of leaves that have not coalesced by epoch i. The proof follows the same lines as the proof of Eq. (3.10). The initial state of the chain is $g_n = (0,0,\ldots,0) \in \mathbb{G}_n^n$ and the final absorbing state is $g_1 = (1,0,0,\ldots,0) \in \mathbb{G}_n^1$. The number of leaf lineages that coalesced from epoch i to epoch $i-1$ is $g_{i,n} - g_{i-1,n}$, where $\mathbb{G}_n^{(i-1)} \ni g_{i-1} \prec_g g_i \in \mathbb{G}_n^{(i)}$. Note that $(g_{i,n} - g_{i-1,n}) \in \{0,1,2\}$, for any $i \in \{2,3,\ldots,n\}$. Out of the $\binom{i}{2}$ many choices for coalescent events among a pair of i lineages during epoch i, only three type of events need to be discriminated in $\mathbb{G}_n$. First, when $(g_{i,n} - g_{i-1,n}) = 0$ we have a coalescent event between two specific non-leaf lineages, each with coalescent vintage smaller than n. Thus, there is exactly one such event among $\binom{i}{2}$ possibilities. Second, when $(g_{i,n} - g_{i-1,n}) = 1$ we have a coalescent event between one specific non-leaf lineage and any one of $g_{i,n}$ many leaf lineages. Thus, there are exactly $\binom{g_{i,n}}{1} = g_{i,n}$ many events among $\binom{i}{2}$ possibilities of the second type. Third, when $(g_{i,n} - g_{i-1,n}) = 2$ we have a coalescent event between any two of $g_{i,n}$ many leaf lineages. Thus, there are exactly $\binom{g_{i,n}}{2}$ many events among $\binom{i}{2}$ possibilities of the third type. All three types of events are accounted for in Eq. (3.15) and thus we have proved Eq. (3.15).

The probability that $g_i \in \mathbb{G}_n^i$ is visited by the chain is obtained by considering the inverse images, $\mathscr{G}^{-1}(g_i)$:

$$
P(g_i) = P(\mathscr{G}^{-1}(g_i)) = \sum_{d_j \in \mathscr{G}^{-1}(g_i)} P(d_j).
$$

with $P(d_j)$ given in Eq. (3.11). The probability $P(g_i) = P(\mathscr{G}^{-1}(g_i))$ can be written explicitly as follows. Let $L = i - g_{i,n}$ be the number of non-leaf lineages in epoch i. Let also $\mathfrak{f}$ be the mapping defined by $\mathfrak{f}(g_i, j_1, \ldots, j_L) = d_i \in \mathbb{D}_n^{(i)}$, where $d_{i,j} = 0$ if and

only if $g_{i,j}=0$, $d_{i,n}=g_{i,n}$ and $d_{i,j}=j_k\geq 2$ if and only if $g_{i,j}$ is the k-th entry which is bigger than zero. In words, the function $\mathfrak{f}$ creates an element $d_i\in\mathbb{D}_n^{(i)}$ compatible with g_i by assigning block sizes $j_1,\ldots,j_L$ to the blocks with vintages less than n (the blocks of vintage n being singletons by construction). The probability $P(g_i)$ is then

$$P(g_i)=\sum_{j_1=2}^{n-g_{i,n}-2(L-1)}\ \sum_{j_2=2}^{n-g_{i,n}-j_1-2(L-2)}\cdots \sum_{j_{L-1}=2}^{n-g_{i,n}-\sum_{i=1}^{L-2}j_i-2} P\left(\mathfrak{f}\left(g_i,j_1,\ldots,j_{L-1},n-g_{i,n}-\sum_{i=1}^{L-1}j_i\right)\right)\ . \quad (5.5)$$

The transition probabilities of the forward jump chain $\{G^{\downarrow}(k)\}_{k\in[n]_+}$ can be obtained from Bayes' rule as follows:

$$P(g_i|g_{i-1})=\begin{cases}P(g_{i-1}|g_i)\frac{P(g_i)}{P(g_{i-1})} & \text{if } g_{i-1}\prec_g g_i\in\mathbb{G}_n^i\\ 0 & \text{otherwise.}\end{cases}.$$

The probability of a g-sequence in Eq. (3.18) can be obtained as follows:

$$P(g)=\prod_{i=n}^{2}P(g_{i-1}|g_i)=\prod_{i=n}^{2}\binom{g_{i,n}}{g_{i,n}-g_{i-1,n}}\binom{i}{2}^{-1}$$
$$=\frac{n!}{2^{\beth(g)}}\prod_{i=2}^{n}\binom{i}{2}^{-1}=\frac{2^{n-\beth(g)-1}}{(n-1)!}. \quad (5.6)$$

The proof is similar to that of Eq. (3.13). Note that Tajima (1983, Eqn. 1) obtains $P(g)$ by another argument.

Proof 5 (of Prop. 8) First we prove that the number of elements in $\mathbb{G}_n^{(i)}$ is, for $i<n$,

$$|\mathbb{G}_n^{(i)}|=\sum_{k=0}^{i-1}\binom{n-i-1}{k}, \quad (5.7)$$

with the convention that $\binom{a}{b}=0$ if $a<b$. For $i=n$, we only have one element, namely a sequence of only 0s, and so $|\mathbb{G}_n^{(n)}|=1$.

Now let $i<n$ and let $g_i\in\mathbb{G}_n^{(i)}$. Since we have i lineages in epoch i, we have at the most i non-zero entries in g_i. In g_i, we have $g_{i,j}=0$ for $j=1,\ldots,i-1$. Further, $g_{i,i}=1$. The remaining $n-1-i$ elements are 0 or 1, with at most $i-1$ 1s. For k non-zero entries in the remaining elements, we have $\binom{n-1-i}{k}$ possibilities to assign the 0s and 1s. Summing over all possible k-values yields Eq. (5.7).

From Eq. (5.7) we have, by summing over all i,

$$|\mathbb{G}_n|=\sum_{i=1}^{n}|\mathbb{G}_n^{(i)}|=\sum_{i=1}^{n-1}\sum_{k=0}^{i-1}\binom{n-i-1}{k}+1.$$

By basic properties of the binomial coefficient, we get,

$$\sum_{i=1}^{n-1}\sum_{k=0}^{i-1}\binom{n-i-1}{k} = \sum_{k=0}^{n-2}\sum_{j=k}^{n-2-k}\binom{j}{k} = \sum_{k=0}^{n-2}\binom{n-k-1}{k+1} = \sum_{k=1}^{n-1}\binom{n-k}{k} = \text{Fibo}(n+1)-1 \quad (5.8)$$

which proves Eq. (3.19).

Proof 6 We derived $P(g)$, the probability of a g-sequence, in Prop. 7. Recall that $|\mathscr{B}_n| = n!(n-1)!\,2^{-(n-1)}$ and that b-sequences are drawn according to the uniform law on $\mathscr{B}_n$. Hence, $P(g) = P(\mathscr{G}\circ\mathscr{D}(\{B^{\uparrow}(t)\}) = g) = |(\mathscr{G}\circ\mathscr{D})^{-1}(g)|/|\mathscr{B}_n|$. This gives us

$$|(\mathscr{G}\circ\mathscr{D})^{-1}(g)| = 2^{-(n-1)}\,n!\,(n-1)!P(g) = 2^{-(n-1)}\,n!\,(n-1)!\,\frac{2^{n-\beth(g)-1}}{(n-1)!} = n!\,2^{-\beth(g)}.$$

Thus, Eq. (3.20) gives us the number of ranked labeled trees that map to any given g-sequence g based on $\beth(g)$, the number of cherries of g. Due to the bijection from $\mathscr{B}_n$ to $\mathscr{C}_n$ and the uniform distribution on $\mathscr{B}_n$ and $\mathscr{C}_n$, the probability $P(c|g) = P(b|g)$

$$P(c|g) = \frac{P(c,g)}{P(g)} = \frac{P(c)}{P(g)} = \frac{P(\mathscr{B}^{-1}(c))}{P(g)} = \frac{P(b)}{P(g)} = \frac{P(b,g)}{P(g)} = P(b|g) \ .$$

Now,

$$P(b|g) = P(c|g) = \frac{P(c)}{P(g)} = \frac{2^{n-1}(n!(n-1)!)^{-1}}{2^{n-\beth(g)-1}((n-1)!)^{-1}} = 2^{\beth(g)}/n! \ .$$

Proof 7 (of Prop. 11) Equations (3.27), (3.28) and (3.29) have been derived by Sainudiin et al (2011). The probability that $\{F^{\uparrow}(k)\}_{k\in[n]_-}$ visits a particular $f_i \in \mathbb{F}_n^i$ at the i-th epoch in Eq. (3.28) has also been given by Tavaré (1983, Equation (7.11)).

Now we prove that the probability of an f-sequence in terms of its shape statistics is given by Eq. (3.30). This probability is given by the product:

$$P(f) = \prod_{i=2}^{n} P(f_i|f_{i-1}) \ . \quad (5.9)$$

For any $f \in \mathscr{F}_n$, we can simplify $P(f)$ given by Eqs. (5.9) and (3.29), as follows:

$$P(f) = \prod_{i=2}^{n} P(f_i|f_{i-i}) = \prod_{i=2}^{n}\left(2^{\mathbb{1}_{\{1\}}\max\langle f_i - f_{i-1}\rangle}\ \ddot{f}_i\ (n-i+1)^{-1}\right)$$
$$= \frac{2^{\sum_{i=2}^{n}\mathbb{1}_{\{1\}}\max\langle f_i - f_{i-1}\rangle}}{(n-1)!}\prod_{i=2}^{n}\ddot{f}_i = \frac{2^{\daleth(f)}}{(n-1)!}\prod_{i=2}^{n}\ddot{f}_i \ .$$

We get Eq. (3.30) from the definition of $\daleth(f)$ in Eq. (3.25) as the number of distinctly-sized lineage splits in f.

Proof 8 (of Prop. 12) The uniform probability on $\mathscr{C}_n$ given by $2^{n-1}(n!(n-1)!)^{-1}$ invokes the probability on f-sequences in $\mathscr{F}_n$ via the inverse image of $\mathscr{F}^{-1}$, i.e.,

$$P(f) = P(\mathscr{F}^{-1}(f)) = |\mathscr{F}^{-1}(f)|\, 2^{n-1}(n!(n-1)!)^{-1}$$

and we have the first equality in Eq. (3.32). The second equality in Eq. (3.32) follows from substituting $P(f)$ by the expression in Eq. (3.30). The probability $P(c|f)$ at Eq. (3.33) follows from

$$P(c|f) = \frac{P(c,f)}{P(f)} = \frac{P(c)}{P(f)} \ .$$

Proof 9 (of Prop. 13) The first equality in Eq. (3.34) is due to the bijection between $\mathscr{D}_n$ and $\mathscr{G}_n$. For the second equality in Eq. (3.34), we establish $|\mathscr{F}'^{-1}(f)| = 2^{-\widehat{\beth}(f)} \prod_{i=2}^{n} \ddot{f}_i$ next. Recall that out of the $n-1$ splits in an f, $\beth(f)$ many of them are cherries and directly lead to leaves while $\daleth(f)$ many of them lead to distinctly-sized splits. The number of remaining splits in f was then defined as $\widehat{\beth}(f) := n-1-\daleth(f)-\beth(f)$.

Let us highlight the following two facts: (1) for any $b, b' \in \mathscr{D}^{-1}(\mathscr{F}'^{-1}(f)) = \mathscr{C}^{-1}(\mathscr{F}^{-1}(f)) \subseteq \mathscr{B}_n$, $P(b) = P(b') = 2^{n-1}(n!(n-1)!)^{-1}$, and (2) for any $d, d' \in \mathscr{F}'^{-1}(f)$ and any $g, g' \in \mathscr{G}(\mathscr{F}'^{-1}(f))$, $P(d) = P(d') = P(g) = P(g') = 2^{n-1-\beth(f)}/(n-1)!$, since $\beth(f) = \sum_{i=2}^{n} f_{i,2} = \beth(d) = \beth(d') = \beth(g) = \beth(g')$. Therefore, the number of ranked tree shapes mapped by a given f-sequence is the number of ranked labeled trees of an f-sequence divided by the number of ranked labeled trees of a g- or d-sequence with the same number of cherries as the f-sequence:

$$\begin{aligned} |\mathscr{F}'^{-1}(f)| = |\mathscr{G}(\mathscr{F}'^{-1}(f))| = \frac{|\mathscr{C}^{-1}(\mathscr{F}^{-1}(f))|}{|\mathscr{D}^{-1}(\mathscr{G}^{-1}(g))|} = \frac{|\mathscr{F}^{-1}(f)|}{n!\,2^{-\beth(g)}} = \frac{|\mathscr{F}^{-1}(f)|}{n!\,2^{-\beth(f)}} \\ = 2^{\daleth(f)+\beth(f)+1-n} \prod_{i=2}^{n} \ddot{f}_i, \end{aligned}$$

where we use Eq. (3.20) for the third-last equality and Eq. (3.32) for the last equality. Finally, Eq. (3.34) follows from the definition of $\widehat{\beth}(f) := n-1-\daleth(f)-\beth(f)$. We get Eq. (3.35) from $P(g)$ in Eq. (3.18), $P(f)$ at Eq. (3.30) and the definition of $\widehat{\beth}(f)$ as follows:

$$\begin{aligned} P(g|f) = \frac{P(g,f)}{P(f)} = \frac{P(g)}{P(f)} = \frac{2^{n-\beth(g)-1}((n-1)!)^{-1}}{2^{\daleth(f)}((n-1)!)^{-1}\prod_{i=2}^{n}\ddot{f}_i} = \frac{2^{n-1-\daleth(f)-\beth(f)}}{\prod_{i=2}^{n}\ddot{f}_i} \\ = 2^{\widehat{\beth}(f)} \left(\prod_{i=2}^{n} \ddot{f}_i\right)^{-1} \ . \end{aligned}$$

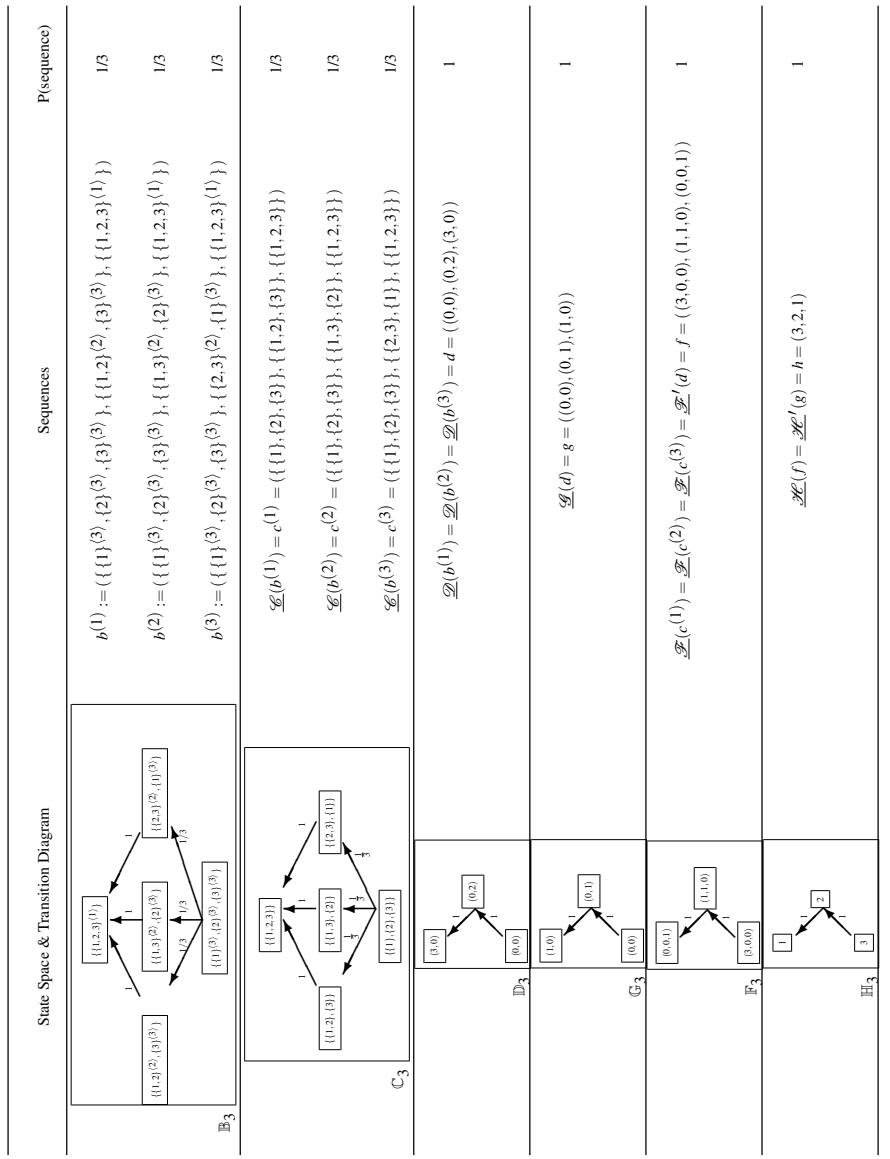


Table 4 When $n = 3$ we tabulate the state spaces, (backward) transition diagrams, the sequences and their probabilities at six resolutions of the n-coalescent.