

# Coalescent experiments I: Unlabeled $n$ -coalescent and the site frequency spectrum

Raazesh Sainudiin<sup>1</sup>, Kevin Thornton<sup>2</sup>, Robert Griffiths<sup>3</sup>, Gil McVean<sup>3</sup> and Peter Donnelly<sup>3</sup>

<sup>1</sup> Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, NZ

<sup>2</sup> Department of Ecology and Evolutionary Biology, University of California, Irvine, USA

<sup>3</sup> Department of Statistics, University of Oxford, Oxford, UK



UCDMS 2009/7 (May 19, 2009). Some rights reserved.

This work is licensed under the Creative Commons Attribution-NonCommercial-Share Alike 3.0 New Zealand Licence. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/nz/>.

**Abstract.** We derive the transition structure of a Markovian lumping of Kingman's  $n$ -coalescent [1, 2]. Lumping a Markov chain is meant in the sense of [3, def. 6.3.1]. The lumped Markov process, referred as the unlabeled  $n$ -coalescent, is a continuous-time Markov chain on the set of all integer partitions of the sample size  $n$ . We derive the backward-transition, forward-transition, state-specific, and sequence-specific probabilities of this chain. We show that the likelihood of any given site-frequency-spectrum (SFS), a commonly used statistics in genome scans, from a locus free of intra-locus recombination, can be directly obtained by integrating conditional realizations of the unlabeled  $n$ -coalescent. We develop a controlled Markov chain for importance sampling such integrals from an augmented unlabeled  $n$ -coalescent forward in time. We apply the methods to population-genetic data to conduct demographic inference at the empirical resolution of the site-frequency-spectra. We also extend a family of classical hypothesis tests of standard neutrality at a non-recombining locus based on any statistics of the SFS to a more powerful version that conditions on the topological information contained in the SFS. We formalize a graph of coalescent experiments to set a decision-theoretic stage for population genetic inference across different empirical resolutions.

**keywords.** partially ordered  $n$ -coalescent experiments graph; controlled Markov chain for importance sampling

# 1 Introduction

Models in population genetics are highly structured stochastic processes [4]. Inference is typically conducted with data that is modeled as a partial observation of one realization of such a process. Likelihood methods are inferentially desirable when they are based on a family of population genetic models for the probability of an observation at the finest empirical resolution available to the experimenter. One typically observes DNA sequences of length  $m$  with a common ancestral history from  $n$  individuals who are currently present in an extant population and uses this information to infer some aspect of the population's history. Unfortunately, it is computationally prohibitive to evaluate the likelihood  $P(u_o|\phi)$  of the data  $u_o \in \mathcal{U}_n^m$  that was observed at the finest available empirical resolution, given a parameter  $\phi \in \mathcal{P}$ , that is indexing a biologically motivated family of models. The sample space of the multiply aligned homologous DNA sequences  $\mathcal{U}_n^m := \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}^{n \times m}$  is doubly indexed by  $n$ , the number of sampled individuals, and  $m$ , the number of sequenced homologous sites. In an ideal world, the optimal inference procedure would be based on the minimally sufficient statistic and implemented in a computing environment free of engineering constraints. Unfortunately, minimally sufficient statistics of data at the currently finest resolution of  $\mathcal{U}_n^m$  are unknown beyond the simplest models of mutation with small values of  $n$  [5–8]. Computationally-intensive inference, based on an observed  $u_o \in \mathcal{U}_n^m$ , with realistically large  $n$  and  $m$ , is currently infeasible for recombining loci and prohibitive for non-recombining loci.

An alternative inference strategy that is computationally feasible involves a relatively low-dimensional statistic  $R(u_o) = r_o$  of the observed *multiple sequence alignment* or *MSA* data  $u_o \in \mathcal{U}_n^m$ . In this approach, one attempts to approximate the likelihood  $P(u_o|\phi)$  or the posterior distribution  $P(\phi|u_o)$ , on the basis of a summary  $r_o = R(u_o)$  of the observed data  $u_o$ , where  $R(u) = r : \mathcal{U}_n^m \rightarrow \mathcal{R}_n^m$  is a statistic with  $\mathcal{R}_n^m$  as its sample space. Since  $R$  is typically not a sufficient statistic for  $\phi$ , i.e.  $P(\phi|r) \neq P(\phi|u)$ , such methods have been termed as *approximate likelihood computations (ALC)* [9] in a frequentist setting or as *approximate Bayesian computations (ABC)* [10, 11] in a Bayesian setting as described in the companion article [12]. Any formal notion of approximate sufficiency for population genetic data  $u_o \in \mathcal{U}_n^m$  must account for the fact that the likelihood  $P(u_o|\phi) = \sum \int_{c_t \in \mathcal{C}_n \mathbb{T}_n} P(u_o|c_t, \phi) P(c_t|\phi)$  is defined by the  $n$ -coalescent prior mixture over elements in a partially observed genealogical space  $\mathcal{C}_n \mathbb{T}_n$  (described in §3.2). This space is both discrete, to account for the sequence of coalescence events, and continuous, to account for the number of generations between such events in units of rescaled time.

The rest of the paper is organized as follows. The basic form of population genetic data and statistics of interest to this paper are briefly introduced in §2. The statistical experiments with  $n$ -coalescents, including the underlying probability models, are introduced in §3. Brief applications in parameter estimation and testing are done in §4.

## 2 Data and Statistics

The data  $u$  is the DNA multiple sequence alignment or MSA obtained from a sample of  $n$  individuals in a population at  $m$  homologous sites. This is assumed to be the finest empirical resolution available to our experimenter. In this paper, we are interested in the posterior distribution over the parameter space  $\mathbb{P}$  on the basis of the observed *site frequency spectrum* or *SFS*  $x_o$ . At a finer resolution we can conduct inference on the basis of the observed *binary incidence matrix* or *BIM*  $v_o$  that is sufficient for Watterson's infinitely-many-sites model of mutation [13] using existing importance sampling methods for models that are more sophisticated than those considered here (e.g. [14–19]). In this study we are not interested in inference at the resolution of BIM and focus instead on SFS. Next, we formalize the BIM and SFS statistics.

We can obtain the site frequency spectrum  $x$  of a given multiply-aligned DNA sequence data  $u$  from a standard encoding of  $u$  into a binary incidence matrix or BIM  $v$ . We describe this encoding next. Let the nucleotide state of the sample's most recent common ancestral sequence be encoded as 0 at all  $m$  sites. Let the derived state at one or more of the samples at a given site be encoded as 1 if it arose from a mutation event in that site's ancestral history. Such bi-allelic sites are common in population genetic data and a site with both an ancestral state and a derived state is called a single nucleotide polymorphism (SNP). Such data is typically modeled, as described in §3, using the infinitely-many-sites model of mutation over the  $n$ -coalescent model of sample genealogy [1, 2]. The binary states at each of the  $m$  homologous sites in  $n$  sampled individuals is denoted by the BIM  $v \in \mathcal{V}_n^m := \{0, 1\}^{n \times m}$ . Let  $v'$  denote the transpose of the binary incidence matrix  $v$  and let  $\mathbb{1}_A(a)$  be the indicator function of some set  $A$  (i.e., if  $a \in A$ , then  $\mathbb{1}_A(a) = 1$ , else  $\mathbb{1}_A(a) = 0$ ). Let the *site sum spectrum* or *SSS*  $w$  corresponding to  $v$  and the corresponding site frequency spectrum or SFS  $x$  be

$$W(v) = w := v' \cdot (1, 1, \dots, 1) : \mathcal{V}_n^m \rightarrow \mathcal{W}_n^m := \{0, 1, 2, \dots, n-1\}^m ,$$

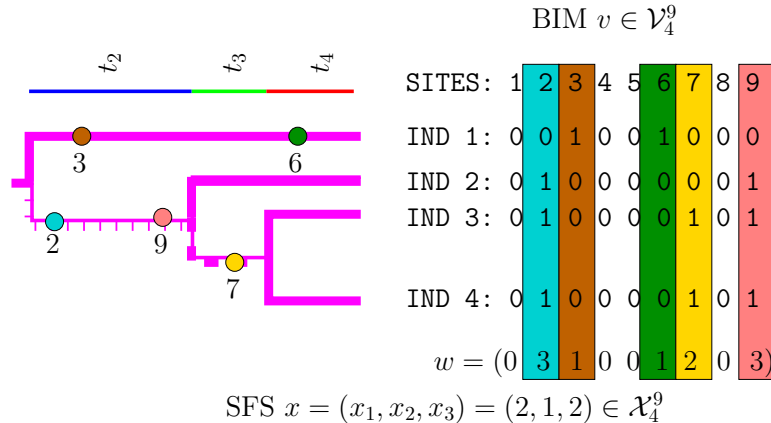
$$X(v) = x := (x_1, \dots, x_{n-1}) \in \mathcal{X}_n^m := \{x \in \mathbb{Z}_+^{n-1} : \sum_{i=1}^{n-1} x_i \leq m\} .$$

One can obtain the SFS  $x$  from a BIM  $v$  via SSS  $w$  as follows:

$$X(v) := X'(W(v)) = x : \mathcal{V}_n^m \rightarrow \mathcal{X}_n^m ,$$

$$X'(w) := \left( \sum_{j=1}^m \mathbb{1}_{\{1\}}(w_j), \sum_{j=1}^m \mathbb{1}_{\{2\}}(w_j), \dots, \sum_{j=1}^m \mathbb{1}_{\{n-1\}}(w_j) \right) = x : \mathcal{W}_n^m \rightarrow \mathcal{X}_n^m .$$

Fig. 1 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. In [12] we show how to obtain various classical statistics of interest from SFS. Next, we describe the basic probability models over  $\mathcal{V}_n^m$  and  $\mathcal{X}_n^m$ .



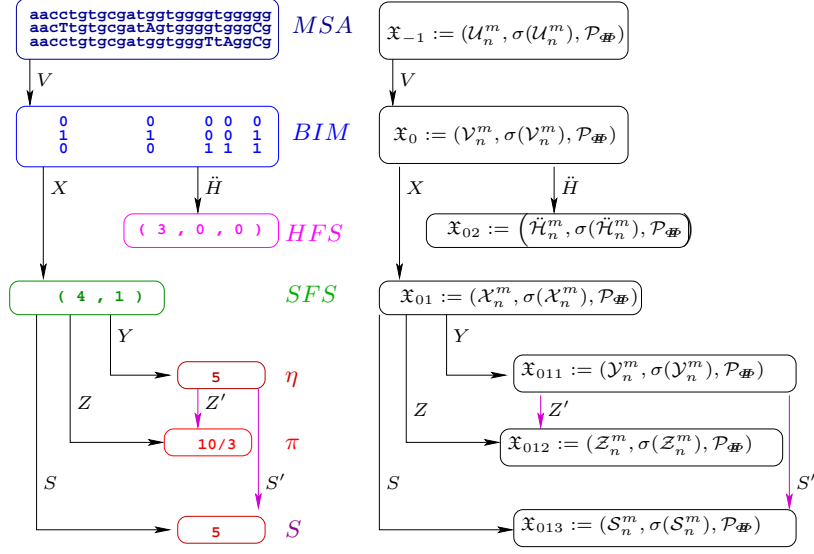
**Fig. 1.** At most one mutation per site under the infinitely-many-sites model are superimposed as a homogeneous Poisson process upon the realization of identical coalescent trees at nine homologous SITES labeled  $\{1, 2, \dots, 9\}$  that constitute a non-recombining locus from four INDividuals labeled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

### 3 An $n$ -Coalescent Experiments Graph

We give the statistical formalities needed to graphically frame our  $n$ -coalescent statistical experiments:  $\mathfrak{X}_{01} := (\mathcal{X}_n^m, \sigma(\mathcal{X}_n^m), \mathcal{P}_{\mathfrak{F}})$  is pursued in this article while  $\mathfrak{X}_{011} := (\mathcal{Y}_n^m, \sigma(\mathcal{Y}_n^m), \mathcal{P}_{\mathfrak{F}})$  and the product of  $\mathfrak{X}_{012} := (\mathcal{Z}_n^m, \sigma(\mathcal{Z}_n^m), \mathcal{P}_{\mathfrak{F}})$  and  $\mathfrak{X}_{013} := (\mathcal{S}_n^m, \sigma(\mathcal{S}_n^m), \mathcal{P}_{\mathfrak{F}})$  are pursued in the companion article [12]. Recall that a statistical experiment  $(\mathcal{X}_n^m, \sigma(\mathcal{X}_n^m), \mathcal{P}_{\mathfrak{F}})$  is the ordered triple consisting of (1) the sample space  $\mathcal{X}_n^m$ , (2) a sigma-algebra over the sample space  $\sigma(\mathcal{X}_n^m)$  and (3) an identifiable  $\mathfrak{F}$ -indexed family of probability measures  $\mathcal{P}_{\mathfrak{F}}$ , i.e.  $\mathfrak{F} \ni \phi \mapsto P_\phi \in \mathcal{P}_{\mathfrak{F}}$ , over the sample space, such that,  $P_\phi := P(x|\phi) \in \mathcal{P}_{\mathfrak{F}}$  for each  $\phi \in \mathfrak{F}$ . Our samples spaces  $\mathcal{V}_n^m$  and  $\mathcal{X}_n^m$  are finite and therefore  $P_\phi$ 's are dominated by the counting measure. Our continuous parameter space in this study is two-dimensional, i.e.  $\mathfrak{F} := (\mathfrak{F}_1, \mathfrak{F}_2) \subset \mathbb{R}_+^2$ . The first parameter  $\phi_1$  is the per-locus mutation rate scaled by the effective population size and is often denoted by  $\theta$  in population genetics literature. The second parameter  $\phi_2$  is the growth rate of our population whose size is growing exponentially from the past. 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.

**Definition 1 (Sufficiency).** A statistic  $T_{\alpha,\beta}(z_\alpha) = z_\beta : \mathcal{Z}_\alpha \rightarrow \mathcal{Z}_\beta$  is sufficient for the experiment  $\mathfrak{X}_\alpha = (\mathcal{Z}_\alpha, \sigma(\mathcal{Z}_\alpha), \mathcal{P}_{\mathfrak{F}})$ , provided:

$$P(Z_\alpha = z_\alpha | T_{\alpha,\beta}(z_\alpha) = z_\beta, \phi) = P(Z_\alpha = z_\alpha | T_{\alpha,\beta}(z_\alpha) = z_\beta),$$



**Fig. 2.** An  $n$ -coalescent experiments graph. An observed multiple sequence alignment of the mother experiment and its offspring are shown on the left. The corresponding formalities are shown on the right.

for any  $\phi \in \Phi$ . This is the classical definition of sufficiency. Given a sufficient statistic  $T_{\alpha,\beta}$  for the experiment  $\mathfrak{X}_\alpha$  and a prior density such that  $P(\phi) \neq 0$  for all  $\phi \in \Phi$ , we get Bayes sufficiency in the Kolmogorov sense [20], in terms of the following posterior identity:

$$P(\phi|z_\alpha) = P(\phi|T_{\alpha,\beta}(z_\alpha) = z_\beta) .$$

**Definition 2 (The Experiments Graph).** Consider an  $\mathfrak{A}$ -indexed set of experiments  $\{\mathfrak{X}_\alpha, \alpha \in \mathfrak{A}\}$ . Let,  $T_{\alpha,\beta} : \mathcal{Z}_\alpha \rightarrow \mathcal{Z}_\beta$ , for some  $\alpha, \beta \in \mathfrak{A}$  with  $\sigma(\mathcal{Z}_\alpha) \supset \sigma(\mathcal{Z}_\beta)$  be a statistic (measurable map). Let  $\mathfrak{M}$  be a set of such maps as well as the identity map. Then, the directed graph of experiments  $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$  with nodes  $\{\mathfrak{X}_\alpha, \alpha \in \mathfrak{A}\}$  and directed edges from a node  $\mathfrak{X}_\alpha$  to a node  $\mathfrak{X}_\beta$ , provided there exists an  $T_{\alpha,\beta} \in \mathfrak{M}$ , is the experiments graph. Consider the partial ordering  $\succ_{\mathfrak{X}}$  induced on the experiments in  $\{\mathfrak{X}_\alpha, \alpha \in \mathfrak{A}\}$  by the maps in  $\mathfrak{M}$ , i.e.,  $\mathfrak{X}_\alpha \succ_{\mathfrak{X}} \mathfrak{X}_\beta$  if and only if there exists a composition of maps from  $\mathfrak{M}$  given by  $T_{\alpha,\beta}^\circ := T_{\alpha,i} \circ T_{i,j} \circ \dots \circ T_{i',j'} \circ T_{j',\beta} : \mathcal{Z}_\alpha \rightarrow \mathcal{Z}_\beta$ , such that  $\sigma(\mathcal{Z}_\alpha) \supset \sigma(\mathcal{Z}_\beta)$ . Then, by construction, (1) the random variables  $\{X_\alpha, \alpha \in \mathfrak{A}\}$  that are adapted to this partially ordered filtration, i.e., for each  $\alpha \in \mathfrak{A}$ ,  $X_\alpha$  is  $\sigma(\mathcal{X}_\alpha)$ -measurable, such that (2)  $E(|X_\alpha|) < \infty$  for all  $\alpha \in \mathfrak{A}$ , form a martingale relative to  $\mathcal{P}_\Phi$  and the partially ordered filtration on  $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$ , i.e.,  $E(X_\alpha | \sigma(X_\beta)) = X_\beta$ , provided  $\mathfrak{X}_\alpha \succ_{\mathfrak{X}} \mathfrak{X}_\beta$ .

In an  $n$ -coalescent experiments graph  $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$  on an  $\mathfrak{A}$ -indexed set of  $n$ -coalescent experiments with a family of statistics  $\mathfrak{M}$ , as partly constructed in §3.2, for instance, there are three distinct linearly ordered sequential asymptotics at every

experiment  $\mathfrak{X}_\alpha$ , in addition to the partially-ordered filtration on  $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$ . This triple asymptotics is a peculiar aspect of the  $n$ -coalescent experiments. The first one involves the sequential limit in the number of sampled individuals  $n \in \mathbb{N}$ , i.e.,  $n \rightarrow \infty$ . The second one involves the sequential limit in the number of sites  $m \in \mathbb{N}$ , i.e.,  $m \rightarrow \infty$ . The first two asymptotics only involve one non-recombining locus of  $m$  DNA sites sampled from  $n$  individuals. The third limit results from a product of single-locus experiments involving the number of sampled loci  $k \in \mathbb{N}$ , i.e.,  $k \rightarrow \infty$ . The product structure is justified under the assumption of infinite recombination between the loci. Thus, asymptotic statistical properties of estimators, for instance, have at least three pure senses of  $\rightarrow \infty$  and several bi/tri-sequential mixed senses of  $(n, m, k) \rightarrow (\infty, \infty, \infty)$  with distinct asymptotic rates of convergence that are of decision-theoretic interest. See [21] and references therein for treatments of the three asymptotics in the pure sense. In the sequel, we are primarily interested in the relative information across different  $n$ -coalescent experiments in our  $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$  for one locus with fixed values of  $n$  and  $m$ . We are not interested in asymptotic experiments, ‘shooting’ out of each node of our experiments graph along the  $n \rightarrow \infty$ ,  $m \rightarrow \infty$ , and/or  $k \rightarrow \infty$  axes, in this paper and instead focus on the ‘small’ or fixed sample experiments in our graph  $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$ . There is only a finite collection of sequentially ordered filtrations, corresponding to the unique paths through  $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$  from the coarsest to the finest empirical resolution. However, in a ‘scientific/technological limit’ one would expect  $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$  itself to grow. It is worth noting that the experiment nodes at the finest resolutions of  $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$  were non-existent over two decades ago, the large values of  $n$ ,  $m$ , and  $k$  one encounters today were non-existent half a decade ago and empirical resolutions that are much finer than our finest resolution of gap-free MSA are readily available today.

The two classes of  $\Phi$ -indexed probability models we consider here are Kingman’s labeled  $n$ -coalescent [1, 2] for the experiment  $(\mathcal{V}_n^m, \sigma(\mathcal{V}_n^m), \mathcal{P}_\Phi)$  with BIM and Kingman’s unlabeled  $n$ -coalescent for the experiment  $(\mathcal{X}_n^m, \sigma(\mathcal{X}_n^m), \mathcal{P}_\Phi)$  with SFS [1, 5.2] (also see [22, p. 136-137] and [23]). As pointed out earlier, inference methods for the BIM experiment  $(\mathcal{V}_n^m, \sigma(\mathcal{V}_n^m), \mathcal{P}_\Phi)$  using importance samplers (e.g. [14–19]) are more developed than those for the SFS experiment  $(\mathcal{X}_n^m, \sigma(\mathcal{X}_n^m), \mathcal{P}_\Phi)$ . In the interest of computationally efficient inference from the coarser resolution of SFS, we develop the probability models in §3.3–3.4 and the associated inference methods in §3.5.

The labeled and unlabeled  $n$ -coalescents approximate the sample genealogy of a non-recombining homologous locus from  $n$  labeled and unlabeled individuals, respectively. The  $n$  samples are randomly drawn from a large exponentially growing Wright-Fisher population [24, 25]. The locus of interest in each sampled individual consists of a DNA sequence of length  $m$  that has undergone selectively neutral mutations under Watterson’s infinitely-many-sites model [13]. The  $n$ -coalescents provide the basic probability models underlying our experiments of interest as they provide a prior mixture over the partially observed genealogical space of coalescent trees  $\mathcal{C}^n \mathbb{T}_n$  (described in §3.2). The other experiment nodes in the experiments graph of Fig. 2 are included to decision-theoretically unify

various classical population genetic experiments. They include  $(\dot{\mathcal{H}}_n^m, \sigma(\dot{\mathcal{H}}_n^m), \mathcal{P}_{\dot{\mathcal{H}}})$  that is based on the haplotype frequency spectrum or HFS  $\dot{H}$  [26, 27], and the three liner sub-experiments of  $(\mathcal{X}_n^m, \sigma(\mathcal{X}_n^m), \mathcal{P}_{\mathcal{X}})$  pursued in the companion article [12]: (i)  $(\mathcal{Y}_n^m, \sigma(\mathcal{Y}_n^m), \mathcal{P}_{\mathcal{Y}})$ , (ii)  $(\mathcal{Z}_n^m, \sigma(\mathcal{Z}_n^m), \mathcal{P}_{\mathcal{Z}})$  and (iii)  $(\mathcal{S}_n^m, \sigma(\mathcal{S}_n^m), \mathcal{P}_{\mathcal{S}})$  that are based on the folded site frequency spectrum or FSFS  $Y$ , the heterozygosity  $Z$ , and the number of segregating sites  $S = \sum_{i=1}^{n-1} x_i$ , respectively.

In this paper and its companion paper [12] we focus on specific experiments. The decision problem of computationally efficient parameter estimation, for instance, on the basis of statistics at a given node in the experiments graph requires an integration over a sufficient equivalence class in  $\mathcal{C}_n \mathbb{T}_n$ , the hidden space of  $n$ -coalescent trees. By further unifying our  $n$ -coalescent models in the hidden space via the theory of lumped  $n$ -coalescent Markov chains [28, §2] we can obtain a lumped  $n$ -coalescent graph [28, §4] that underpins the unified multi-resolution  $n$ -coalescent of [28]. Through this lumped  $n$ -coalescent graph, the companion structure in the hidden space of our  $n$ -coalescent experiments graph  $\mathfrak{G}_{\mathfrak{X}, \mathfrak{Y}}$ , it is also possible to take decisions that fully exploit the partially ordered filtrations that are indexed by sub-graphs of  $\mathfrak{G}_{\mathfrak{X}, \mathfrak{Y}}$ .

### 3.1 Number of Ancestral Lineages of a Wright-Fisher Sample

In the simple Wright-Fisher discrete generation model with a constant population size  $N$ , i.e. the exponential growth rate  $\phi_2 = 0$ , the offspring “choose” their parents uniformly and independently at random from the previous generation due to the uniform multinomial sampling of  $N$  offspring from the  $N$  parents in the previous generation. First, note that the following ratio can be approximated:

$$\begin{aligned} \frac{N_{[j]}}{N^j} &:= \frac{N}{N} \frac{N-1}{N} \cdots \frac{N-(j-1)}{N} = 1 \left(1 - \frac{1}{N}\right) \cdots \left(1 - \frac{j-1}{N}\right) = \prod_{k=1}^{j-1} (1 - kN^{-1}) \\ &= 1 - N^{-1} \sum_{k=1}^{j-1} k + O(N^{-2}) = 1 - \binom{j}{2} N^{-1} + O(N^{-2}) \quad . \end{aligned}$$

Let  $S_i^{(j)}$  denote the Stirling number of the second kind, i.e.  $S_i^{(j)}$  is the number of set partitions of a set of size  $i$  into  $j$  blocks. Thus, the  $N$ -specific probability of  $i$  extant sample lineages in the current generation becoming  $j$  extant lineages in the previous generation is:

$${}^N P_{i,j} = \begin{cases} S_i^{(i)} (N_{[i]} N^{-i}) = 1 (N_{[i]} N^{-i}) = 1 - \binom{i}{2} N^{-1} + O(N^{-2}) & : \text{if } j = i \\ S_i^{(i-1)} (N_{[i-1]} N^{-i}) = \binom{i}{2} (N^{-1} N_{[i-1]} N^{-(i-1)}) = \\ \binom{i}{2} N^{-1} (1 - N^{-1} \binom{i-1}{2}) + O(N^{-2}) = \binom{i}{2} N^{-1} + O(N^{-2}) & : \text{if } j = i - 1 \\ S_i^{(i-\ell)} (N_{[i-\ell]} N^{-i}) = S_i^{(i-\ell)} (N^{-\ell} N_{[i-1]} N^{-(i-\ell)}) = \\ S_i^{(i-\ell)} N^{-\ell} (1 - N^{-1} \binom{i-\ell}{2}) + O(N^{-2}) = O(N^{-2}) & : \text{if } j = i - \ell, \\ 0 & : \text{otherwise} \quad , \end{cases} \quad (1)$$

where,  $1 < \ell < i - 1$ .

In words, the probability that any specific pair of lineages, among the  $\binom{i}{2}$  many pairs of the currently extant  $i$  ancestors of the  $n$  sampled lineages, coalesces in one generation is  $1/N$  and that this pair remains distinct for more than  $g$  generations is  $(1 - 1/N)^g$ . Let  $\mathbb{Z}_- := \{0, -1, -2, \dots\}$  denote an ordered and countably infinite discrete time index set. Next, we rescale time in this discrete time Markov chain  $\{^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$  over the state space  $\mathbb{H}_n := \{n, n-1, \dots, 1\}$  with 1-step transition probabilities (1) termed *the death chain of the number of ancestral sample lineages within the Wright-Fisher population of constant size  $N$* . Let the rescaled time  $t$  be  $g$  in units of  $N$  generations. Then, the probability that a pair of lineages remain distinct for more than  $t$  units of the rescaled time is:  $(1 - 1/N)^{\lfloor Nt \rfloor} \xrightarrow{N \rightarrow \infty} e^{-t}$ .

The transition probabilities  $P_{i,j}(t)$  of the *pure death process*  $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$ , in the rescaled time  $t$  over the state space  $\mathbb{H}_n$ , is a limiting continuous time Markov chain approximation of the  $\lfloor Nt \rfloor$ -step transition probabilities  $^N P_{i,j}(\lfloor Nt \rfloor)$  of the discrete time death chain with 1-step transition probabilities (1), as the population size  $N$  tends to infinity:

$$^N P_{i,j}(\lfloor Nt \rfloor) \xrightarrow{N \rightarrow \infty} P_{i,j}(t) = \exp(Qt), \quad \text{where, } q_{i,i-1} = \binom{i}{2}, q_{i,i} = -\binom{i}{2},$$

$q_{i,j} = 0$  for all other  $(i, j) \in \mathbb{H}_n \times \mathbb{H}_n$  but with 1 as an absorbing state. The matrix  $Q$  is called the instantaneous rate matrix of the death process Markov chain  $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$  and its  $(i, j)$ -th entry is  $q_{i,j}$ . Thus, the  $i$ -th epoch-time random variable  $T_i$  during which time there are  $i$  distinct ancestral lineages of our sample is approximately exponentially distributed with rate parameter  $\binom{i}{2}$  and is independent of other epoch-times. In other words, for large  $N$ , the random vector  $T = (T_2, T_3, \dots, T_n)$  of epoch-times, corresponding to the transition times of the pure death process  $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$  on the state space  $\mathbb{H}_n$ , has the product exponential density  $\bigotimes_{i=2}^n \binom{i}{2} e^{-\binom{i}{2} t_i}$  over its support  $\mathbb{T}_n := \mathbb{R}_+^{n-1}$ . Note that the initial state of  $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$  is  $n$ , the final absorbing state is 1 and the embedded jump chain  $\{H^\uparrow(k)\}_{k \in [n]_-}$  of this death process, termed *the embedded death chain*, deterministically marches from  $n$  to 1 in decrements of 1 over  $\mathbb{H}_n$ , where,  $[n]_- := \{n, n-1, \dots, 2, 1\}$  denotes an ordered discrete time index set.

### 3.2 Kingman's Labeled $n$ -Coalescent

Next, we model the sample genealogy at a finer resolution than the number of ancestral lineages of our Wright-Fisher sample of size  $n$ . If we assign distinct labels to our  $n$  samples and want to trace the ancestral history of these sample-labeled lineages then Kingman's labeled  $n$ -coalescent lends a helping hand. Let  $\mathbb{C}_n$  be the set of all set partitions of the label set  $\mathcal{L} = \{1, 2, \dots, n\}$  of our  $n$  samples. Denote by  $\mathbb{C}_n^{(i)}$  the set of all set partitions with  $i$  blocks, i.e.,  $\mathbb{C}_n = \bigcup_{i=1}^n \mathbb{C}_n^{(i)}$ . Let  $c_i := (c_{i,1}, c_{i,2}, \dots, c_{i,i}) \in \mathbb{C}_n^{(i)}$  denote the  $i$  elements of  $c_i$  in canonical order and let  $|c_{i'}|$  denote the number of elements in  $c_{i'} \in \mathbb{C}_n^{(|c_{i'}|)} \in \mathbb{C}_n$ . The *labeled  $n$ -coalescent partial ordering*  $\prec_c$  on  $\mathbb{C}_n$  is based on the immediate

precedence relation  $\prec_c$ :

$$c_{i'} \prec_c c_i \iff c_{i'} = c_i \setminus c_{i,j} \setminus c_{i,k} \cup (c_{i,j} \cup c_{i,k}), j \neq k, j, k \in \{1, 2, \dots, |c_i|\}.$$

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 any distinct pair of elements in  $c_i$ . Thus,  $c_{i'} \prec_c c_i$  implies  $|c_{i'}| = |c_i| - 1$ .

Consider the discrete time Markov chain  $\{C^\uparrow(k)\}_{k \in [n]_-}$  on  $\mathbb{C}_n$  with initial state  $C^\uparrow(n) = c_n = \{\{1\}, \{2\}, \dots, \{n\}\}$  and final absorbing state  $C^\uparrow(1) = c_1 = \{\{1, 2, \dots, n\}\}$ , with the following transition probabilities [2, (2.2)]:

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

Now, let  $c := (c_n, c_{n-1}, \dots, c_1)$  be a  $c$ -sequence or coalescent sequence obtained from the sequence of states visited by a realization of the chain, and denote the space of such  $c$ -sequences by

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

The probability that  $c_i \in \mathbb{C}_n^{(i)}$  is visited by the chain [2, (2.3)] is:

$$P(c_i) = \frac{(n-i)! i! (i-1)!}{n! (n-1)!} \prod_{j=1}^i |c_{i,j}|! , \quad (3)$$

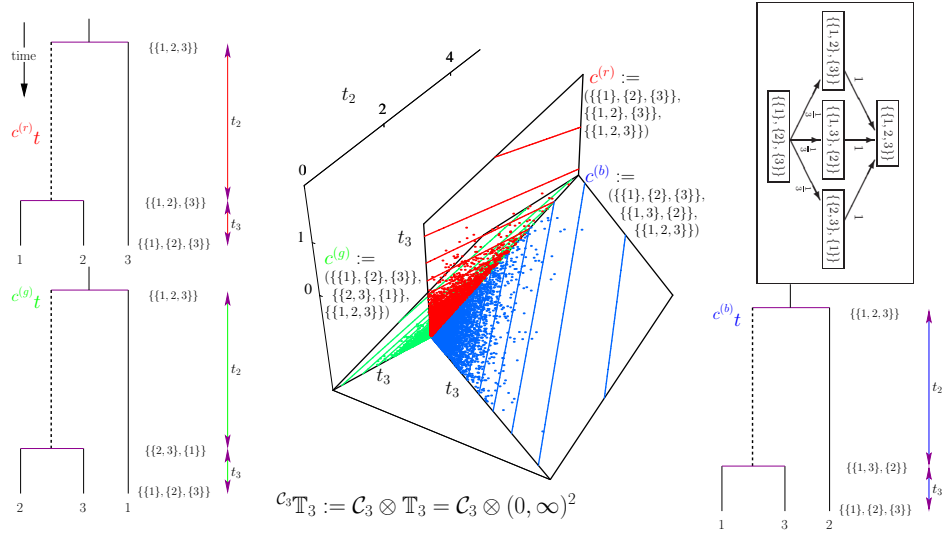
and the probability of a  $c$ -sequence is uniformly distributed over  $\mathcal{C}_n$  with

$$P(c) = \prod_{i=n}^2 P(c_{i-1}|c_i) = \frac{2^{n-1}}{n! (n-1)!} = \frac{1}{|\mathcal{C}_n|} . \quad (4)$$

Kingman's labeled  $n$ -coalescent [1, 2] is a continuous-time Markov chain  $\{C^\uparrow(t)\}_{t \in \mathbb{R}_+}$  on  $\mathbb{C}_n$  with rate matrix  $Q$ . The entries  $q(c_{i'}|c_i)$ ,  $c_i, c_{i'} \in \mathbb{C}_n$  of  $Q$ , specifying the transition rate from state  $c_i$  to  $c_{i'}$ , are [1, (2.10)]:

$$q(c_{i'}|c_i) = \begin{cases} -\binom{i}{2} & : \text{if } c_i = c_{i'}, c_i \in \mathbb{C}_n^{(i)} \\ 1 & : \text{if } c_{i'} \prec_c c_i \\ 0 & : \text{otherwise .} \end{cases} \quad (5)$$

The above instantaneous transition rates for the continuous time Markov chain  $\{C^\uparrow(t)\}_{t \in \mathbb{R}_+}$  are obtained by coupling the independent death process  $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$  of §3.1 over  $\mathbb{H}_n$  with the discrete time Markov chain  $\{C^\uparrow(k)\}_{k \in [n]_-}$  on  $\mathbb{C}_n$ . This continuous time Markov chain approximates the appropriate  $N$ -specific discrete time Markov chain over  $\mathbb{C}_n$  that is modeling the ancestral genealogical history of a sample of size  $n$  labeled by  $\mathcal{L}$  and taken at random from the Wright-Fisher population of constant size  $N$ . This asymptotic approximation, as the population size  $N \rightarrow \infty$ , can be seen using arguments similar to those in §3.1. See [2, (§1-2)] for this construction.



**Fig. 3.** Realizations of 3-coalescent trees in the space of such trees is plotted on the three rectangles as colored points in middle panel. The lines on the rectangles are the contours of the independent exponentially distributed epoch times for each  $c$ -sequence. Each of the three coalescent trees, with two branch lengths  $(t_3, t_2)$ , representing a realization in the corresponding rectangle and the transition probability diagram of the the Markov chain  $\{C^t(k)\}_{k \in \{3,2,1\}}$  on  $\mathcal{C}_3$  are shown counter clock-wise in the four corner panels, respectively.

Let the space of *ranked, rooted, binary, phylogenetic trees* with leaves or samples labeled by  $\mathcal{L} = \{1, 2, \dots, n\}$  [29, §2.3] further endowed with branch or lineage lengths under a *molecular clock* — i.e. the lineage length obtained by summing the epoch-times from each sample (labeled leaf) to the root node or *the most recent common ancestor* (MRCA) is the same — be constructively defined by the  $n$ -coalescent as:

$$\mathcal{C}_n \mathbb{T}_n := \mathcal{C}_n \otimes \mathbb{T}_n := \{c t := (c^n t_n, c^{n-1} t_{n-1}, \dots, c^2 t_2) : c \in \mathcal{C}_n, t \in \mathbb{T}_n := \mathbb{R}_+^{n-1}\}.$$

$\mathcal{C}_n \mathbb{T}_n$  is called the  $n$ -coalescent tree space. An  $n$ -coalescent tree  $c t \in \mathcal{C}_n \mathbb{T}_n$  describes the ancestral history of the sampled individuals. Fig. 3 depicts the  $n$ -coalescent tree space  $\mathcal{C}_3 \mathbb{T}_3$  for the sample label set  $\mathcal{L} = \{1, 2, 3\}$  with sample size  $n = 3$ .

### 3.3 Kingman's Unlabeled $n$ -Coalescent

The unlabeled  $n$ -coalescent is mentioned as a lumped Markov chain of the labeled  $n$ -coalescent and termed the ‘label-killed’ process by Kingman [1, 5.2]. Tavaré [22, p. 136-137] terms it the ‘family-size process’ along the nomenclature of a more general birth-death-immigration process [23]. The transition probabilities

of this Markov process, in either temporal direction, are not explicitly developed in [1] or [22]. They are developed here along with the state and sequence-specific probabilities.

Consider the coalescent epoch at which there are  $i$  lineages. Let  $f_{i,j}$  denote the number of lineages subtending  $j$  leaves, i.e. the frequency of lineages that are ancestral to  $j$  samples, at this epoch. Let us summarize these frequencies from the  $i$  lineages as  $j$  varies over its support by  $f_i := (f_{i,1}, f_{i,2}, \dots, f_{i,n})$ . Then the space of  $f_i$ 's is defined by,

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

Let the set of such frequencies over all epochs be  $\mathbb{F}_n := \bigcup_{i=1}^n \mathbb{F}_n^{(i)}$ . Note that  $\mathbb{F}_n$  contains the frequency of the cardinalities of sets belonging to every element of  $\mathbb{C}_n$ , the state space of Kingman's labeled  $n$ -coalescent. Thus,  $\mathbb{F}_n$  is the frequency representation of the integer partitions of  $n$ , i.e. the solutions to the Diophantine equation  $\{(p_1, p_2, \dots, p_n) \in \mathbb{Z}_+^n : \sum_{i=1}^n i p_i = n\}$ , and  $\mathbb{F}_n^{(i)}$  are those integer partitions of  $n$  composed of  $i$  positive integers. Let us define an  $f$ -sequence  $f$  as:

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

where,  $\prec_f$  is the immediate precedence relation that induces the partial ordering  $\preceq_f$  on  $\mathbb{F}_n$ . It is defined by denoting the  $j$ -th unit vector of length  $n$  by  $e_j$ , as follows:

$$f_{i'} \prec_f f_i \iff f_{i'} = f_i - e_j - e_k + e_{j+k}. \quad (6)$$

Thus,  $\mathcal{F}_n$  is the space of  $f$ -sequences with  $n$  samples. One can see  $\mathcal{F}_n$  as the space of the frequencies of the cardinalities of  $c$ -sequences in  $\mathbb{C}_n$ . Recall the  $c$ -sequence  $c = (c_n, c_{n-1}, \dots, c_1)$ , where  $c_{i-1} \prec_c c_i$ ,  $c_{i-1} \in \mathbb{C}_n^{i-1}$ ,  $c_i \in \mathbb{C}_n^i$ , and  $c_i := (c_{i,1}, c_{i,2}, \dots, c_{i,i})$  contains its canonically ordered  $i$  subsets. Then the corresponding  $f$ -sequence is given by the map  $\underline{\mathcal{F}}(c) = f : \mathbb{C}_n \rightarrow \mathcal{F}_n$ , as follows:

$$\underline{\mathcal{F}}(c) := (\mathcal{F}(c_n), \dots, \mathcal{F}(c_1)), \mathcal{F}(c_i) := \left( \sum_{h=1}^i \mathbf{1}_{\{1\}}(|c_{i,h}|), \dots, \sum_{h=1}^i \mathbf{1}_{\{n\}}(|c_{i,h}|) \right). \quad (7)$$

Note that  $\mathcal{F}_n$  indexes an equivalence class in  $\mathbb{C}_n$  via  $\underline{\mathcal{F}}^{[-1]}(f)$ , the inverse map of (7). Having defined  $f$ -sequences and their associated spaces, we define a discrete time Markov chain  $\{F^\uparrow(k)\}_{k \in [n]_-}$  on  $\mathbb{F}_n$  that is analogous to the Markov chain  $\{C^\uparrow(k)\}_{k \in [n]_-}$  on  $\mathbb{C}_n$  given by (2). This is the embedded discrete time Markov chain of the unlabeled  $n$ -coalescent.

**Proposition 1 (Backward Transition Probabilities of an  $f$ -sequence)**

The probability of  $f := (f_n, f_{n-1}, \dots, f_1) \in \mathcal{F}_n$  under the  $n$ -coalescent is given by the product:

$$P(f) = \prod_{i=n}^2 P(f_{i-1}|f_i), \quad (8)$$

such that  $P(f_{i-1}|f_i)$  are the backward transition probabilities of a Markov chain  $\{F^\uparrow(k)\}_{k \in [n]_-}$  on  $\mathbb{F}_n$ , with  $f_i \in \mathbb{F}_n^{(i)}$ ,  $f_{i-1} \in \mathbb{F}_n^{(i-1)}$ :

$$P(f_{i-1}|f_i) = \begin{cases} f_{i,j}f_{i,k} \binom{i}{2}^{-1} & : \text{if } f_{i-1} = f_i - e_j - e_k + e_{j+k}, j \neq k \\ \binom{f_{i,j}}{2} \binom{i}{2}^{-1} & : \text{if } f_{i-1} = f_i - e_j - e_k + e_{j+k}, j = k \\ 0 & : \text{otherwise} \end{cases} \quad (9)$$

where, the initial state is  $f_n = (n, 0, \dots, 0)$  and the final absorbing state is  $f_1 = (0, 0, \dots, 1)$ .

*Proof.* Since (8) is merely a consequence of (9), we prove (9) next. When there are  $i$  lineages in Kingman's labeled  $n$ -coalescent, a coalescence event can reduce the number of lineages to  $i - 1$  by coalescing one of  $\binom{i}{2}$  many pairs. Hence, the inverse  $\binom{i}{2}^{-1}$  appears in the transition probabilities. Out of these pairs, there are two kinds of pairs that need to be differentiated. The first type of coalescence events involve pairs of edges that subtend the same number of leaves. Since  $f_{i,j}$  many edges subtend  $j$  leaves, there are  $\binom{f_{i,j}}{2}$  many pairs that lead to this event (case when  $j = k$ ). The second type of coalescence events involve pairs of edges that subtend different number of leaves. For any distinct  $j$  and  $k$ ,  $f_{i,j}f_{i,k}$  many pairs would lead to coalescence events between edges that subtend  $j$  and  $k$  leaves (case when  $j \neq k$ ). Note that our condition that  $f_{i-1} = f_i - e_j - e_k + e_{j+k}$  for each  $i \in \{n, n-1, \dots, 3, 2\}$  ensures that our  $f$  remains in  $\mathcal{F}_n$  as we go backwards in time from the  $n$ -th coalescent epoch with  $n$  samples to the first one with the single ancestral lineage.

**Proposition 2 (Probability of an  $f_i$ )**

The probability that the Markov chain  $\{F^\uparrow(k)\}_{k \in [n]_-}$  visits a particular  $f_i \in \mathbb{F}_n^{(i)}$  at the  $i$ -th epoch (given in [22, Equation (7.11)]) is:

$$P(f_i) = \frac{i!}{\prod_{j=1}^i f_{i,j}!} \binom{n-1}{i-1}^{-1} \quad (10)$$

*Proof.* Recall that  $f_{i,j}$  is the number of edges during the  $i$ -th coalescent epoch (during which there are  $i$  lineages) that subtend  $j$  leaves, where,  $j \in \{1, 2, \dots, n\}$ , i.e.  $\sum_{j=1}^n f_{i,j} = i$  and  $\sum_{j=1}^n j f_{i,j} = n$ . Now, label the  $i$  edges in some arbitrary manner. Let the number of the subtended leaves from the  $i$  labeled edges be  $\Lambda := (\Lambda_1, \Lambda_2, \dots, \Lambda_i)$ . Due to the  $n$ -coalescent,  $\Lambda$  is a random variable with a uniform distribution on integer partitions of  $n$ , such that  $\sum_{j=1}^i \Lambda_j = n$  and  $\Lambda_j \geq 1$ . Thus,  $P(\Lambda) = \binom{n-1}{i-1}^{-1}$ . Since there are  $i!/\prod_{j=1}^i f_{i,j}!$  many ways of labeling the  $i$  edges, we get the  $P(f_i)$  as stated.

**Proposition 3 (Forward Transition Probabilities of an  $f$ -sequence)**

The probability of  $f := (f_n, f_{n-1}, \dots, f_1) \in \mathcal{F}_n$  is given by the product:

$$P(f) = \prod_{i=2}^n P(f_i|f_{i-1}), \quad (11)$$

such that  $P(f_i|f_{i-1})$  are the forward transition probabilities of a Markov chain  $\{F^\downarrow(k)\}_{k \in [n]_+}$  on  $\mathbb{F}_n$  with the ordered time index set  $[n]_+ := \{1, 2, \dots, n\}$ :

$$P(f_i|f_{i-1}) = \begin{cases} 2f_{i-1,j+k}(n-i+1)^{-1} & : \text{if } f_i = f_{i-1} + e_j + e_k - e_{j+k}, j \neq k, \\ & j+k > 1, f_i \in \mathbb{F}_n^{(i)}, f_{i-1} \in \mathbb{F}_n^{(i-1)} \\ f_{i-1,j+k}(n-i+1)^{-1} & : \text{if } f_i = f_{i-1} + e_j + e_k - e_{j+k}, j = k, \\ & j+k > 1, f_i \in \mathbb{F}_n^{(i)}, f_{i-1} \in \mathbb{F}_n^{(i-1)} \\ 0 & : \text{otherwise} \end{cases} \quad (12)$$

with initial state  $f_1 = (0, 0, \dots, 1)$  and final absorbing state  $f_n = (n, 0, \dots, 0)$ .

Note that we canonically write a sequential realization  $(f_1, f_2, \dots, f_n)$  of  $\{F^\downarrow(k)\}_{k \in [n]_+}$  in reverse order as the  $f$ -sequence  $f = (f_n, f_{n-1}, \dots, f_1)$ .

*Proof.* Since (11) is merely a consequence of (12), we prove (12) next. An application of the definition of conditional probability twice, followed by Proposition 2 yields:

$$\begin{aligned} P(f_i|f_{i-1}) &= P(f_{i-1}|f_i)P(f_i)/P(f_{i-1}) \\ &= P(f_{i-1}|f_i) \frac{i!}{\prod_{h=1}^i f_{i,h}!} \binom{n-1}{i-1}^{-1} / \frac{(i-1)!}{\prod_{h=1}^{i-1} f_{i-1,h}!} \binom{n-1}{i-2}^{-1} \\ &= P(f_{i-1}|f_i) \frac{\prod_{h=1}^{i-1} f_{i-1,h}!}{\prod_{h=1}^i f_{i,h}!} \frac{i(i-1)}{n-(i-1)} \end{aligned}$$

Next we substitute  $P(f_{i-1}|f_i)$  of Proposition 1 for the first case:  $f_i = f_{i-1} + e_j + e_k - e_{j+k}, j \neq k, j+k > 1$ , i.e. the coordinates of  $f_i$  and  $f_{i-1}$  are such that  $f_{i,j} = f_{i-1,j} + 1, f_{i,k} = f_{i-1,k} + 1, f_{i,j+k} = f_{i-1,j+k} - 1$ , and  $f_{i,h} = f_{i-1,h}, \forall h \in \{1, 2, \dots, n\} \setminus \{j, k, j+k\}$ .

$$\begin{aligned} P(f_i|f_{i-1}) &= f_{i,j}f_{i,k} \binom{i}{2}^{-1} \frac{\prod_{h=1}^{i-1} f_{i-1,h}!}{\prod_{h=1}^i f_{i,h}!} \frac{i(i-1)}{n-(i-1)} \\ &= f_{i,j}f_{i,k} \frac{f_{i-1,j}!f_{i-1,k}!f_{i-1,j+k}!}{f_{i,j}!f_{i,k}!f_{i,j+k}!} \frac{2}{n-(i-1)} \\ &= f_{i,j}f_{i,k} \frac{(f_{i,j}-1)!(f_{i,k}-1)!(f_{i,j+k}+1)!}{f_{i,j}!f_{i,k}!f_{i,j+k}!} \frac{2}{n-(i-1)} \\ &= \frac{2(f_{i,j+k}+1)}{n-(i-1)} = 2f_{i-1,j+k}(n-i+1)^{-1} \end{aligned}$$

A similar substitution of  $P(f_{i-1}|f_i)$  of Proposition 1 for the second case:  $f_i = f_{i-1} + e_j + e_k - e_{j+k}, j = k, j+k > 1$ , i.e.,  $f_{i,j} = f_{i-1,j} + 2, f_{i,2j} = f_{i-1,2j} - 1$

and  $f_{i,h} = f_{i-1,h}, \forall h \in \{1, 2, \dots, n\} \setminus \{j, 2j\}$ .

$$\begin{aligned}
P(f_i|f_{i-1}) &= \binom{f_{i,j}}{2} \binom{i}{2}^{-1} \frac{\prod_{h=1}^{i-1} f_{i-1,h}!}{\prod_{h=1}^i f_{i,h}!} \frac{i(i-1)}{n-(i-1)} \\
&= \frac{f_{i,j}(f_{i,j}-1)}{n-(i-1)} \frac{f_{i-1,j}! f_{i-1,2j}!}{f_{i,j}! f_{i,2j}!} \\
&= \frac{f_{i,j}(f_{i,j}-1)}{n-(i-1)} \frac{(f_{i,j}-2)!(f_{i,2j}+1)!}{f_{i,j}! f_{i,2j}!} \\
&= \frac{(f_{i,2j}+1)}{n-(i-1)} = f_{i-1,2j}(n-i+1)^{-1} = f_{i-1,j+k}(n-i+1)^{-1}.
\end{aligned}$$

This concludes the proof.

*Kingman's unlabeled  $n$ -coalescent* or the *unvintaged and sized  $n$ -coalescent* in the descriptive nomenclature of [28] is the continuous time Markov chain  $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$  on  $\mathbb{F}_n$  whose rate matrix  $Q = q(f_{i'}|f_i)$  for any two states  $f_i, f_{i'} \in \mathbb{F}_n$  is:

$$q(f_{i'}|f_i) = \begin{cases} -i(i-1)/2 & : \text{if } \mathbb{F}_n^{(i)} \ni f_i = f_{i'}, \\ f_{i,j} f_{i,k} & : \text{if } \mathbb{F}_n^{(i-1)} \ni f_{i'} = f_i - e_j - e_k + e_{j+k}, j \neq k, f_i \in \mathbb{F}_n^{(i)}, \\ (f_{i,j})(f_{i,j}-1)/2 & : \text{if } \mathbb{F}_n^{(i-1)} \ni f_{i'} = f_i - e_j - e_k + e_{j+k}, j = k, f_i \in \mathbb{F}_n^{(i)}, \\ 0 & : \text{otherwise} \end{cases} \quad (13)$$

The initial state is  $f_n = (n, 0, 0, \dots, 0)$  and the final absorbing state is  $f_1 = (0, 0, \dots, 1)$ . The above rates for the continuous time Markov chain  $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$  on  $\mathbb{F}_n$  are obtained by coupling the independent death process  $\{H^\uparrow(t)\}_{t \in \mathbb{R}_+}$  of §3.1 over  $\mathbb{H}_n$  with the discrete time Markov chain  $\{F^\uparrow(k)\}_{k \in [n]_-}$  on  $\mathbb{C}_n$ .

Let  $\{^N F^\uparrow(k)\}_{k \in \mathbb{Z}_-}$  be the discrete time sample genealogical Markov chain of  $n$  unlabeled samples taken at random from the present generation of a Wright-Fisher population of constant size  $N$  over the state space  $\mathbb{F}_n$  analogous to the death chain  $\{^N H^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ . The next Proposition (proved in [28, Prop. 3.28] using the theory of lumped Markov chains) states how  $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$  approximates  $\{^N F^\uparrow(k)\}_{k \in \mathbb{Z}_-}$  on  $\mathbb{F}_n$ .

**Proposition 4 (Kingman's Unlabeled  $n$ -coalescent)**

The  $\lfloor Nt \rfloor$ -step transition probabilities,  $^N P_{f_i, f_{i'}}(\lfloor Nt \rfloor)$ , of the chain  $\{^N F^\uparrow(k)\}_{k \in \mathbb{Z}_-}$ , converge to the transition probabilities of the continuous-time Markov chain  $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$  with rate matrix  $Q$  of (13), i.e.

$$^N P_{f_i, f_{i'}}(\lfloor Nt \rfloor) \xrightarrow{N \rightarrow \infty} P_{f_i, f_{i'}}(t) = \exp(Qt) .$$

*Proof.* See [28, proof of Prop. 3.28].

*Remark 1 (Markovian lumping from  $\mathbb{C}_n$  to  $\mathbb{F}_n$  via  $\mathcal{F}$ ).* Our lumping of Kingman's labeled  $n$ -coalescent over  $\mathbb{C}_n$  to Kingman's unlabeled  $n$ -coalescent over  $\mathbb{F}_n$ , via the mapping  $\mathcal{F}$ , is Markov as pointed out by Kingman [1, (5.1), (5.2)]

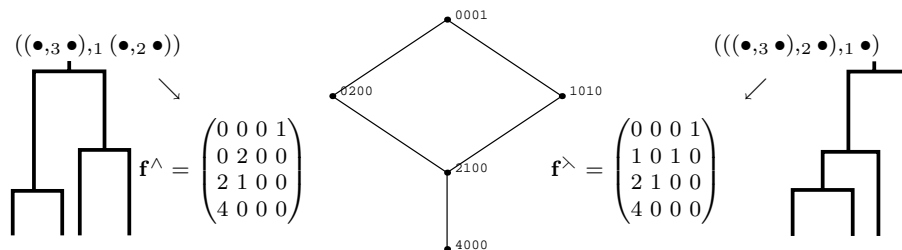
using the arguments in [30, Sec. IIIId]. See [28, §2.1] for an introduction to lumped processes and [28, proof of Prop. 3.29] for a proof that  $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$  is a Markov lumping of  $\{C^\uparrow(t)\}_{t \in \mathbb{R}_+}$ .

Next we provide some concrete examples of  $c$ -sequences and their lumping into  $f$ -sequences for small  $n$ . When there are 2 samples there is one  $c$ -sequence  $c = (\{\{1\}, \{2\}\}, \{\{1, 2\}\})$  and one  $f$ -sequence  $f = \underline{\mathcal{F}}(c) = ((2, 0), (0, 1))$ .

**Example 5 (3 Samples)** *When there are 3 samples we have three  $c$ -sequences:  $c^{(r)}$ ,  $c^{(b)}$  and  $c^{(g)}$  (see Fig. 3) and all of them map to the only  $f$ -sequence  $f$ :*

$$\begin{aligned} f &= ((3, 0, 0), (1, 1, 0), (0, 0, 1)) \\ &= \underline{\mathcal{F}}(c^{(r)}) := \underline{\mathcal{F}}(\left(\{\{1\}, \{2\}, \{3\}\}, \{\{1, 2\}, \{3\}\}, \{\{1, 2, 3\}\}\right)) \\ &= \underline{\mathcal{F}}(c^{(b)}) := \underline{\mathcal{F}}(\left(\{\{1\}, \{2\}, \{3\}\}, \{\{1, 3\}, \{2\}\}, \{\{1, 2, 3\}\}\right)) \\ &= \underline{\mathcal{F}}(c^{(g)}) := \underline{\mathcal{F}}(\left(\{\{1\}, \{2\}, \{3\}\}, \{\{2, 3\}, \{1\}\}, \{\{1, 2, 3\}\}\right)) \end{aligned}$$

**Example 6 (4 Samples)** *When there are 3 samples we have two  $f$ -sequences and eighteen  $c$ -sequences. We denote the  $f$ -sequences by  $f^\wedge$  and  $f^\lambda$ . We can apply Equation (7) to  $\mathcal{C}_4$  and find that 12  $c$ -sequences map to  $f^\wedge$  and 6 map to  $f^\lambda$ . They are depicted in Fig. 4.*

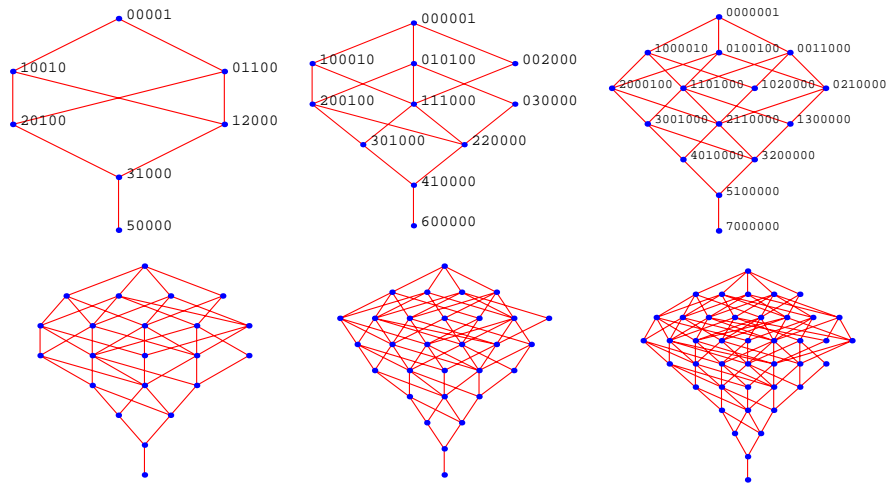


**Fig. 4.** The two  $f$ -sequences  $f^\wedge$  and  $f^\lambda$  corresponding to the balanced (left panel) and unbalanced unlabeled genealogies of four samples (right panel) and the Hasse diagram of the state transition diagrams of  $\{F^\uparrow(k)\}_{k \in [n]_-}$  and  $\{F^\downarrow(k)\}_{k \in [n]_+}$  on  $\mathbb{F}_4$  (middle panel).

Kemeny & Snell [3, 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 we exploit in the sequel where we show that it suffices to analyze the unlabeled  $n$ -coalescent to prescribe measures over  $\mathcal{X}_n^m$ , the sample space of SFS. By lumping the states we are doing far fewer summations during the integration of probabilities over the hidden space of  $f$ -sequences, as opposed to  $c$ -sequences, when evaluating the

likelihood of the observed SFS. The extent of this lumping as  $|\mathbb{F}_n|/|\mathbb{C}_n|$ , the ratio of the number of integer partitions of  $n$  and the  $n$ -th Bell number for a range of sample sizes is tabulated below.

$n =  \mathbb{H}_n $	4	10	30	60	90	120
$ \mathbb{C}_n $	15	$1.2 \times 10^5$	$8.5 \times 10^{23}$	$9.8 \times 10^{59}$	$1.4 \times 10^{101}$	$5.1 \times 10^{145}$
$ \mathbb{F}_n $	5	42	$5.6 \times 10^3$	$9.7 \times 10^5$	$5.7 \times 10^7$	$1.8 \times 10^9$
$ \mathbb{F}_n / \mathbb{C}_n $	0.33	$3.6 \times 10^{-4}$	$6.6 \times 10^{-21}$	$9.9 \times 10^{-55}$	$4.0 \times 10^{-94}$	$3.6 \times 10^{-137}$



**Fig. 5.** Hasse diagrams of the state transition diagrams of the backward and forward Markov chains,  $\{F^\uparrow(k)\}_{k \in [n]_-}$  and  $\{F^\downarrow(k)\}_{k \in [n]_+}$ , respectively, on  $\mathbb{F}_n$  for  $n = 5, 6, 7$  on top row with labeled states and  $n = 8, 9, 10$  in bottom row.

In the Hasse diagram of  $\mathbb{F}_n$ , the states  $f_1, \dots, f_n$  in  $\mathbb{F}_n$  form the nodes or vertices and there is an edge between  $f_i$  and  $f_j$  if  $f_i \prec_f f_j$ , i.e.  $f_i$  immediately precedes  $f_j$ . Each Hasse diagram of  $\mathbb{F}_n$  embodies two directed and weighted graphs of the state transition diagrams of  $\{F^\uparrow(k)\}_{k \in [n]_-}$  and  $\{F^\downarrow(k)\}_{k \in [n]_+}$ . These two state transition graphs are temporally oriented, directed and edge-weighted by the one-step transition probabilities of  $\{F^\uparrow(k)\}_{k \in [n]_-}$  and  $\{F^\downarrow(k)\}_{k \in [n]_+}$ . This structure is used during inference based on SFS data.

Standard graph algorithms may be readily used on the state transition graphs of the unlabeled  $n$ -coalescent (see Fig. 5 for small values of  $n$ ). For instance, Dijkstra's search can be used to compute the  $f$ -sequence with the smallest or largest probability under  $\{F^\uparrow(k)\}_{k \in [n]_-}$  or  $\{F^\downarrow(k)\}_{k \in [n]_+}$  or dynamic programs of stochastic network flow algorithms may be adapted for importance sampling. Using the *Boost Graph Library* [31] it is possible to solve such problems for

larger values of  $n$ . When the nodes in each  $\mathbb{F}_n^{(i)}$  are in reverse-lexicographic order (as done in Fig. 5), the least probable and the most probable  $f$ -sequences tend to hover the left and right edges of the graphs in Fig. 5, respectively (See [28, Prop. 3.31]), as  $n$  gets large.

### 3.4 Exponentially Growing Population

Thus far, we have focused on stochastic processes whose realizations yield labeled and unlabeled sample genealogies of a Wright-Fisher population of constant size  $N$ . Consider a demographic model of steady exponential growth forward in time:

$$N(t) = N(0) \exp(\phi_2 t),$$

where  $N(0)$  is the current population size. One can apply a deterministic time-change to the epoch times of the constant population model to obtain the epoch times of the growing population [22]:

$$P\left(T_k > t \mid \sum_{j=k+1}^n T_j = t_{k+1:n}\right) = \exp\left(-\binom{k}{2} \phi_2^{-1} \exp(\phi_2 t_{k+1:n}) (\exp(\phi_2 t) - 1)\right).$$

### 3.5 Inference under the Unlabeled $n$ -Coalescent

Now, we introduce the infinitely-many-sites model [13] of mutations that are super-imposed on the labeled and unlabeled  $n$ -coalescent sample genealogies in order to prescribe the  $\phi$ -indexed measures on our sample spaces  $\mathcal{V}_n^m$  and  $\mathcal{X}_n^m$ , respectively. Recall that a coalescent tree  ${}^c t$  realized under the  $n$ -coalescent describes the labeled ancestral history of the sampled individuals as a binary tree. Fig. 1 shows a coalescent tree for a sample of four individuals. In neutral models considered here under parameter  $\phi = (\phi_1, \phi_2) \in \mathfrak{P}$ , mutations are super-imposed upon the coalescent trees at each site according to a homogeneous Poisson process at rate  $\phi_1 l_\bullet$ , where  $\phi_1 := 4N_e \mu m$ ,  $l_\bullet$  is the total size of the tree,  $N_e$  is the effective population size,  $\mu$  is the mutation rate per generation per site. Under the infinitely-many-sites mutation model [13] we further stipulate that at most one mutation is allowed per site. The only parameter in the simplest  $n$ -coalescent model with mutations just described is the scaled per-locus mutation rate  $\phi_1$  for the locus of  $m$  sites where we have assumed the same mutation rate  $\mu$  at all  $m$  sites. The population's exponential growth rate is  $\phi_2$ . For a given coalescent tree  ${}^c t \in \mathcal{C}_n \mathbb{T}_n$ , let the map:

$$L({}^c t) = l := (l_1, l_2, \dots, l_{n-1}) : \mathcal{C}_n \mathbb{T}_n \rightarrow \mathcal{L}_n := \mathbb{R}_+^{n-1} \quad (14)$$

compress the tree  ${}^c t$  into the  $n-1$  lineage lengths that lead to singleton, doubleton,  $\dots$ , and “ $(n-1)$ -ton” mutations, respectively, i.e.  $l_i$  is the length of all the lineages in  ${}^c t$  that subtend  $i$  samples or leaves. For example in Fig. 1, (i) the bold lineage of the tree with label set  $\mathfrak{L} = \{1, 2, 3, 4\}$  upon which the mutations at

sites 3 and 6 occur, lead to singleton mutations, (ii) the bold-dashed lineage upon which the mutation at site 7 occurs leads to doubleton mutations and (iii) the thin-dashed lineage upon which mutations at sites 2 and 9 occur lead to tripleton mutations. Thus,  $l_1$ ,  $l_2$  and  $l_3$  are the lengths of these three types of lineages, respectively. Finally,  $l_\bullet := \sum_{i=1}^{n-1} l_i \in \mathbb{R}_+$  is the total length of all the lineages of the tree  ${}^c t$  that are ancestral to the sample since the most recent common ancestor across all the sites. Now, let  $\bar{l}_i := l_i/l_\bullet$  be the relative length of lineages that subtend  $i$  leaves across the sites. Now, define  $\bar{l} := (\bar{l}_1, \bar{l}_2, \dots, \bar{l}_{n-1}) \in \Delta_{n-2}$ , the  $(n-2)$ -unit-simplex containing all  $\bar{l} \in \mathbb{R}_+^{n-1}$  such that  $\sum_{i=1}^{n-1} \bar{l}_i = 1$ . Then, if  $L({}^c t) = l$ , the following conditional probability of  $x$  is given by the Poisson-Multinomial distribution:

$$P(x|\phi, {}^c t) = P(x|\phi, l) = e^{-\phi_1 l_\bullet} (\phi_1 l_\bullet)^s \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i! , \quad (15)$$

where,  $s = \sum_{i=1}^{n-1} x_i$  is the number of segregating sites. The distribution on  ${}^c \mathbb{T}_n$  is given by the  $\phi_2$ -indexed  $n$ -coalescent approximation of the sample genealogy in an exponentially growing Wright-Fisher model. This distribution on  ${}^c \mathbb{T}_n$  in turn determines the distribution of the random vector  $L$  on  $\mathcal{L}_n$ .

Since  $x$  is a summary of the data  $v$  one could try to apply the ABC/ALC algorithm [12, Alg. 1] to obtain  $P(\phi|x_o)$ , the posterior conditional on the observed  $x_o$ . A direct application of ABC/ALC algorithm however, under any non-trivial metric on  $\mathcal{X}_n^m$  to specify  $\mathbf{m}$ , will be highly inefficient. Moreover, for large  $n$ , a simulated labeled  $n$ -coalescent tree  ${}^c t$  with  $l = L({}^c t)$  will satisfy  $P(x_o|\phi, l) = 0$  with high probability, if one were to simulate  ${}^c t$  directly from a parameter  $\phi_2$  and superimpose mutations upon it according to the parameter  $\phi_1$  after having drawn  $\phi := (\phi_1, \phi_2)$  according to the prior density  $P(\phi)$ .

We remedy this problem of inferring  $\phi$  based on the observed SFS  $x_o$  via a naive ABC/ALC Algorithm [12, Alg. 1] by taking the observations of Kemeny & Snell [3, p. 124] into consideration and employing the appropriate lumped Markov process to efficiently obtain  $P(\phi|x_o)$ . Using the unlabeled  $n$ -coalescent we can directly prescribe the  $\phi$ -indexed family of measures over  $\mathcal{X}_n^m$  and obtain the sampling distribution over  $\mathcal{X}_n^m$ , i.e. the probability of an SFS  $x \in \mathcal{X}_n^m$  when conditioned on the parameter  $\phi$  and an  $f$ -sequence  $f \in \mathcal{F}_n$ . Recall  $P(x|\phi, {}^c t) = P(x|\phi, l)$ ,  $l = L({}^c t)$ , as in (15). We show that  $l$  is determined by the  $f$ -sequence  $f = \mathcal{F}(c)$  of the  $c$ -sequence  $c$  and the epoch-times  $t$  in the coalescent tree  ${}^c t$ . First, we introduce a matrix form  $\mathbf{f}$  of  $f$ . Any  $f$ -sequence  $f = (f_n, f_{n-1}, \dots, f_1)$ , that is a sequential realization under  $\{F^\uparrow(k)\}_{k \in [n]_-}$  or a reverse-ordered sequential realization under  $\{F^\downarrow(k)\}_{k \in [n]_+}$ , can also be written as an  $(n-1) \times (n-1)$  matrix  $\mathbf{F}(f) = \mathbf{f}$  as follows:

$$\mathbf{F} : \mathcal{F}_n \rightarrow \mathbb{Z}_+^{(n-1) \times (n-1)}, \quad \mathbf{F}(f) = \mathbf{f} := \begin{pmatrix} f_{2,1} & f_{2,2} & \cdots & f_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n-1,1} & f_{n-1,2} & \cdots & f_{n-1,n-1} \\ f_{n,1} & f_{n,2} & \cdots & f_{n,n-1} \end{pmatrix} . \quad (16)$$

Thus, the matrix form of  $f = (f_n, f_{n-1}, \dots, f_1)$  or the  $f$ -matrix is the  $(n-1) \times (n-1)$  matrix  $\mathbf{f}$  whose  $(i-1)$ -th row is  $(f_{i,1}, f_{i,2}, \dots, f_{i,n-1})$ , where,  $i = 2, 3, \dots, n$ .

**Proposition 7 (Probability of SFS given  $f$ -sequence and epoch-times)**

Let  ${}^c t \in \mathcal{C}_n \mathbb{T}$  be a given coalescent tree,  $c$  be its  $c$ -sequence,  $f = \mathcal{F}(c)$  be its  $f$ -sequence,  $\mathbf{f} = \mathbf{F}(f)$  be its  $f$ -matrix and  $t = (t_2, t_3, \dots, t_n) \in (0, \infty)^{n-1}$  be its epoch times as a column vector and its transpose  $t^\mathbf{T}$  be the corresponding row vector. Then,  $L({}^c t) = l$  of (14) is given by the following matrix multiplication:

$$l = t^\mathbf{T} \mathbf{f} = \left( \sum_{i=2}^n t_i f_{i,1}, \sum_{i=2}^{n-1} t_i f_{i,2}, \dots, \sum_{i=2}^2 t_i f_{i,n-1} \right) \quad (17)$$

More succinctly,  $l_j = \sum_{i=2}^{n+1-j} t_i f_{i,j}$  for  $j = 1, 2, \dots, n-1$ . And the probability of an SFS  $x$  given a vector of epoch-times  $t \in (0, \infty)^{n-1}$  and any coalescent tree  ${}^c t \in \mathcal{F}^{-1}(f)t := \{{}^c t : c \in \mathcal{F}^{-1}(f)\}$  is:

$$P(x|\phi, {}^c t) = P(x|\phi, l) = P(x|\phi, t^\mathbf{T} \mathbf{f}) = e^{-\phi_1 l_\bullet} (\phi_1 l_\bullet)^s \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i! , \quad (18)$$

where,  $s = \sum_{i=1}^{n-1} x_i$  is the number of segregating sites.

*Proof.* The proof of (17) is merely a consequence of the encoding of  $f$  as the matrix  $\mathbf{f}$  and (18) follows from (17) and (15).

The computation of  $l$  from  $t$  and  $\mathbf{f}$  requires at most  $n^2 - 2n + 1$  multiplications and summations over  $\mathbb{R}$ . Exploiting the predictable sparseness of  $\mathbf{f}$  is more efficient especially for large  $n$ . Thus, given the parameter  $\phi = (\phi_1, \phi_2)$  and a sample size  $n$ , we can efficiently draw SFS samples from  $\mathcal{X}_n^m$  via Algorithm 1.

---

**Algorithm 1** SFS Sampler under Kingman's unlabeled  $n$ -coalescent

---

1: **input:**

1. scaled mutation rate  $\phi_1$  of the locus
2. sample size  $n$

2: **output:** an SFS sample  $x$  from the standard neutral  $n$ -coalescent

3: generate an  $f$ -sequence  $f$  either under  $\{F^\uparrow(k)\}_{k \in [n]_-}$  or  $\{F^\downarrow(k)\}_{k \in [n]_+}$

4: draw  $t \sim T = (T_2, T_3, \dots, T_n) \sim \bigotimes_{i=2}^n \binom{i}{2} e^{-\binom{i}{2} t_i}$ , or as desired

5:  $l = t^\mathbf{T} \cdot \mathbf{f}$ , where  $\mathbf{f} = \mathbf{F}(f)$

6: draw  $x$  from Poisson-Multinomial distribution  $e^{-\phi_1 l_\bullet} (\phi_1 l_\bullet)^s \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!$

7: **return:**  $x$

---

Note that the only restriction on  $t$  is that it be a positive real vector. Thus, any indexed family of measures over  $(0, \infty)^{n-1}$ , including nonparametric ones,

may be used in the computation of  $l = t^T \cdot \mathbf{f}$ , provided the  $c$ -sequence  $c$  and its  $f$ -sequence  $f = \underline{\mathcal{F}}(c)$  are drawn from the labeled  $n$ -coalescent and the corresponding unlabeled  $n$ -coalescent, respectively, in an exchangeable manner that is independent of the epoch-times  $t$ . Furthermore, one may use Algorithm 1 to adapt Algorithm 1 in [12] and conduct approximate Likelihood or Bayesian computations on the basis of summaries that are further compressions of the directly simulable site frequency spectrum as done in [12] via Markov bases.

Next we study one  $f$ -sequence in detail as it is an interesting extreme case that will resurface in the sequel. Let the  $f$ -sequence  $f^\lambda \in \mathcal{F}_n$  denote that of the fully unbalanced tree. Its probability based on (8) and (9) are:

$$f^\lambda := (f_1^\lambda, f_2^\lambda, \dots, f_n^\lambda), \text{ where, } f_i^\lambda = (i-1)e_1 + e_{(n-i+1)}, \quad (19)$$

$$P(f^\lambda) = \prod_{i=n}^2 P(f_{i-1}^\lambda | f_i^\lambda) = \prod_{i=n-1}^2 \frac{(i-1)1}{i(i-1)/2} = \frac{2^{n-2}}{(n-1)!}. \quad (20)$$

The number of  $c$ -sequences corresponding to it is  $|\underline{\mathcal{F}}^{-1}(f^\lambda)| = n!/2$ .

The posterior distribution  $P(\phi|x) \propto P(x|\phi)P(\phi)$  over  $\Phi$  is the object of inferential interest. For an efficient inference based on SFS  $x$ , we first investigate the topological information about the tree  ${}^c t$  that the SFS  $x$  was realized upon. We are only interested in this information provided by the drawn  $x$  and thus can only resolve the topology of  ${}^c t$  up to equivalence classes of  $\underline{\mathcal{F}}^{-1}(f)$ , where  $f$  is the  $f$ -sequence corresponding to the  $c$ -sequence of  ${}^c t$ . For samples of size  $n \leq 3$  there is only one  $f$ -sequence in  $\mathcal{F}_n$ . For samples with  $n \geq 4$ , consider the following mapping of the SFS  $x \in \mathcal{X}_n^m$  into vertices of the unit hyper-cube  $\{0, 1\}^{n-1}$ , a binary encoding of  $\mathbf{2}^{\{1, 2, \dots, n-1\}}$ , the power set of  $\{1, 2, \dots, n-1\}$ :

$$X^\otimes(x) = x^\otimes := (x_1^\otimes, \dots, x_{n-1}^\otimes) := (\mathbf{1}_{\mathbb{N}}(x_1), \dots, \mathbf{1}_{\mathbb{N}}(x_{n-1})) : \mathcal{X}_n^m \rightarrow \{0, 1\}^{n-1}.$$

If  $x_h^\otimes = 1$  then there is at least one mutation in the  $h$ -th entry of the SFS  $x$ , i.e.  $x_h > 0$ . Thus,  $X^\otimes(x) = x^\otimes$  encodes the indices of  $x$  with at least one mutation. Next, consider the following two sets of  $f$ -sequences

$$F_n(x^\otimes) := \bigcup_{\{h: x_h^\otimes=1\}} \{f \in \mathcal{F}_n : \sum_{i=1}^n f_{i,h} = 0\}, \quad \mathcal{C}F_n(x^\otimes) := \mathcal{F}_n \setminus F_n(x^\otimes). \quad (21)$$

The set of  $f$ -sequences  $F_n(x^\otimes)$  and its complement  $\mathcal{C}F_n(x^\otimes)$  play a fundamental role in inference from an SFS  $x$  and its  $X^\otimes = x^\otimes$ . Note that when an SFS  $x$  has none of the  $x_i$ 's equaling 0, then its  $x^\otimes = (1, 1, \dots, 1)$  and  $\mathcal{C}F_n(x^\otimes)$  only contains the  $f$ -sequence corresponding to the completely unbalanced tree  $f^\lambda$  given by (19). At the other extreme, when an SFS  $x$  has all its  $x_i$ 's equaling 0 with  $x^\otimes = (0, 0, \dots, 0)$ , we are unable to discriminate among  $f$ -sequences since  $\mathcal{C}F_n(x^\otimes) = \mathcal{F}_n$ . Thus,

$$\mathcal{C}F_n(0, 0, \dots, 0) = \mathcal{F}_n \quad \text{and} \quad \mathcal{C}F_n(1, 1, \dots, 1) = \{f^\lambda\}. \quad (22)$$

Therefore, the size of  $\mathcal{C}F_n(x^\otimes)$  can range from 1 to  $|\mathcal{F}_n|$ , depending on  $x^\otimes$ . More generally, we have the following Proposition.

**Proposition 8 (Likelihood of SFS)**

For any  $t \in (0, \infty)^{n-1}$  and any  $x \in \mathcal{X}_n^m$  with  $x^\otimes = X^\otimes(x)$ ,

$$f \in F_n(x^\otimes), l = t^T \cdot \mathbf{F}(f) \implies \prod_{i=1}^{n-1} \bar{l}_i^{x_i} = 0, \quad (23)$$

$$P(x|\phi) = \frac{\sum_{f \in \mathcal{C}F_n(x^\otimes)} P(f) \left( \int_{t \in (0, \infty)^{n-1}} \left( e^{-\phi_1 l_\bullet} (\phi_1 l_\bullet)^s \prod_{i=1}^{n-1} \bar{l}_i^{x_i} \right) P(t|\phi) \right)}{\prod_{i=1}^{n-1} x_i!}. \quad (24)$$

*Proof.* We first prove the implication in (23). Given any  $t \in (0, \infty)^{n-1}$  and any  $x \in \mathcal{X}_n^m$  with  $x^\otimes = X^\otimes(x)$ , let  $f \in F_n(x^\otimes)$ . First, suppose  $x_h^\otimes = 0$  for every  $h \in \{1, 2, \dots, n-1\}$ , then  $F_n(x^\otimes) = \emptyset$  and we have nothing to prove. Now, suppose there exists some  $h$  such that  $x_h^\otimes = 1$ , or equivalently  $x_h > 0$ , then by the constructive definition of  $F_n(x^\otimes)$ , we have that for any  $f \in F_n(x^\otimes)$   $\sum_{i=1}^n f_{i,h} = 0$ , which implies that  $f_{i,h} = 0$  for every  $i \in \{1, 2, \dots, n\}$  since  $f_{i,j} \geq 0$ . Therefore, by applying this implication to the expression for  $l_h$  in Proposition 7, we have that  $l_h = \sum_{i=2}^{n+1-h} t_i f_{i,h} = 0$  and finally the desired equality that  $\prod_{i=1}^{n-1} \bar{l}_i^{x_i} = 0$  in (23) is a consequence of  $\bar{l}_h^{x_h} = (l_h/l_\bullet)^{x_h} = 0^{x_h} = 0$ .

Next we prove (24). Repeated application of the definition of conditional probability and the neutral structure of the  $n$ -coalescent model leads to the following expression for  $P(x, \phi)$  in  $P(x|\phi) = P(x, \phi)/P(\phi)$ .

$$\begin{aligned} P(x, \phi) &= \sum_{c \in \mathcal{C}_n} \int_{t \in (0, \infty)^{n-1}} P(x, \phi, t, c) = \sum_{f \in \mathcal{F}_n} \int_{t \in (0, \infty)^{n-1}} P(x, \phi, t, f) \\ &= \sum_{f \in \mathcal{F}_n} \int_{t \in (0, \infty)^{n-1}} P(x|\phi, t, f) P(\phi, t, f) \\ &= \sum_{f \in \mathcal{F}_n} P(f) \int_{t \in (0, \infty)^{n-1}} P(x|\phi, l = t^T \cdot \mathbf{F}(f)) P(t|\phi) P(\phi) \end{aligned}$$

since, by independence of  $f$  and  $(\phi, t)$

$$P(\phi, t, f) = P(f|\phi, t) P(\phi, t) = P(f) P(\phi, t) = P(f) P(t|\phi) P(\phi) .$$

Thus, by letting  $\mathbf{F}(f) = \mathbf{f}$ , the likelihood of the SFS  $x$  is:

$$P(x|\phi) = P(x, \phi)/P(\phi) = \sum_{f \in \mathcal{F}_n} P(f) \int_{t \in (0, \infty)^{n-1}} P(x|\phi, l = t^T \cdot \mathbf{f}) P(t|\phi) .$$

Substituting for  $P(x|\phi, l = t^T \cdot \mathbf{f})$  from Proposition 7 and only summing over  $f \in \mathcal{C}F_n(x^\otimes)$  with non-zero probability  $P(x|\phi, l = t^T \cdot \mathbf{f})$ , we get the discrete sum weighted by integrals on  $\mathbb{T}_n := (0, \infty)^{n-1}$ , the required equality in (24).

Next we devise an algorithm to estimate  $P(x|\phi)$ , the probability of an observed SFS  $x$  given a parameter  $\phi$ . This is accomplished by constructing a

Markov chain  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$  on the state space  $\mathbb{F}_n^{x^\otimes} \subset \mathbb{F}_n \times \{0, 1\}^{n-1}$  such that every sequence of states visited by this chain yields a probable  $f$ -sequences  $f$  for the observed SFS  $x$ , i.e.  $f \in \mathcal{C}F_n(x^\otimes)$ . Obtaining an optimal importance sampler by using the sequential realizations of  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$  and its continuous time variant as a proposal distribution in order to get the Monte Carlo estimate of  $P(x|\phi)$  is necessary and possible. However, this is a subsequent problem in variance reduction of the Monte Carlo estimate for large values of  $n$  that depends further on the precise nature of  $\phi$ -indexed measures on  ${}^c_n\mathbb{T}_n$ . In this paper, we focus on small  $n \in \{4, 5, \dots, 10\}$  and exhaustively sum over all  $f \in \mathcal{C}F_n(x^\otimes)$  that are unique sequential realizations of  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$ . This  $x^\otimes$ -indexed family of  $2^{n-1}$  Markov chains  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$  over state spaces contained in  $\mathbb{F}_n \times \{0, 1\}^{n-1}$  may also be thought of as a controlled Markov chain (e.g. [32, §7.3]) over the state space  $\mathbb{F}_n$  with control space  $\{0, 1\}^{n-1}$  that can produce the desired  $f$ -sequences in  $\mathcal{C}F_n(x^\otimes)$ .

**Proposition 9 (A Proposal over  $\mathcal{C}F_n(x^\otimes)$ )**

For a given SFS  $x \in \mathcal{X}_n^m$  and  $X^\otimes(x) = x^\otimes \in \{0, 1\}^{n-1}$ , consider the discrete time Markov chain  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$  over the state space of ordered pairs  $(f_{i'}, z_{i'}) \in \mathbb{F}_n^{x^\otimes} \subset \mathbb{F}_n \times \{0, 1\}^{n-1}$ , with the initial state given by  $(f_1, x^\otimes) = ((0, 0, \dots, 1), x^\otimes)$ , the transition probabilities obtained by a controlled reweighing of the the transition probabilities of  $\{F^\downarrow(k)\}_{k \in [n]_+}$  over  $\mathbb{F}_n$  as follows:

$$P((f_{i'}, z_{i'})|(f_i, z_i)) = \begin{cases} P(f_{i'}|f_i)/\Sigma(f_i, z_i) & : \text{if } (f_i, z_i) \prec_{f,z} (f_{i'}, z_{i'}) , \\ 0 & : \text{otherwise} , \end{cases} \quad (25)$$

where,

$$\Sigma(f_i, z_i) = \sum_{(j,k) \in \Xi(f_i, z_i)} P(f_i - e_{j+k} + e_j + e_k | f_i),$$

$$\Xi(f_i, z_i) := \{(j, k) : f_{i, j+k} > 0, 1 \leq j \leq \hat{j} \leq k \leq j + k - 1\},$$

$$\hat{j} := \max\{\min\{\max\{\ell : z_{i, \ell} = 1\}, j + k - 1\}, \lceil \frac{j+k}{2} \rceil\},$$

$$(f_i, z_i) \prec_{f,z} (f_{i'}, z_{i'}) \iff \begin{cases} f_{i'} = f_i + e_j + e_k - e_{j+k}, (j, k) \in \Xi(f_i, z_i), \text{ and} \\ z_{i'} = z_i - \mathbf{1}_{\{1\}}(z_{i, j}) e_j - \mathbf{1}_{\{1\}}(z_{i, k}) e_k , \end{cases}$$

and with  $(f_n, (0, 0, \dots, 0)) = ((n, 0, \dots, 0), (0, 0, \dots, 0))$  as the final absorbing state.

Let  $\mathcal{F}_n^{x^\otimes}$  be the set of sequential realizations of the first component of the ordered pairs of states visited by  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$ , i.e.

$$\mathcal{F}_n^{x^\otimes} := \{f = (f_n, f_{n-1}, \dots, f_1) : f_i \in \mathbb{F}_n^{(i)}, (f_i, z_i) \prec_{f,z} (f_{i+1}, z_{i+1}), z_1 = x^\otimes\} .$$

Then  $\mathcal{F}_n^{x^\otimes} = \mathcal{C}F_n(x^\otimes)$ .

*Proof.* We will prove that  $\mathcal{F}_n^{x^\otimes} = \mathbb{C}_{F_n}(x^\otimes)$  for three cases after noting that the ortho-normal basis vector  $e_i$  in  $\{0, 1\}^{n-1}$  and  $\mathbb{F}_n$  takes the appropriate dimension. The first two cases involve constructive proofs.

Case 1: Suppose  $x^\otimes = (0, 0, \dots, 0)$ . Since  $\mathbb{C}_{F_n}(x^\otimes) = \mathcal{F}_n$  by (22), we need to show that  $\mathcal{F}_n^{x^\otimes} = \mathcal{F}_n$ . Initially, at time step 1,

$$F^{\downarrow x^\otimes}(1) = (f_1, z_1) = (f_1, x^\otimes) = ((0, 0, \dots, 0, 1), (0, 0, \dots, 0))$$

Note that for any time step  $i$ ,  $z_i$  in the current state  $(f_i, z_i)$  remains at  $(0, 0, \dots, 0)$ . Thus,  $\max\{\ell : z_{i,\ell} = 1\} = \max\{\emptyset\} = -\infty$  and therefore,

$$\hat{j} := \max\{\min\{\max\{\ell : z_{i,\ell} = 1\}, j + k - 1\}, \lceil \frac{j+k}{2} \rceil\} = \lceil \frac{j+k}{2} \rceil, \text{ and}$$

$$\Xi(f_i, z_i) := \{(j, k) : f_{i,j+k} > 0, 1 \leq j \leq \lceil \frac{j+k}{2} \rceil \leq k \leq j + k - 1\}.$$

Therefore, the first component of the chain can reach all states in  $\mathbb{F}_n$  that are immediately preceded by  $f_i$  under  $\prec_f$  making  $\Sigma(f_i, z_i) = 1$ . Thus, when  $x^\otimes = (0, 0, \dots, 0)$  our fully uncontrolled Markov chain  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$  visits states in  $\mathbb{F}_n$  in a manner identical to the the Markov chain  $\{F^{\downarrow}(k)\}_{k \in [n]_+}$  over  $\mathbb{F}_n$ . Therefore,  $\mathcal{F}_n^{x^\otimes} = \mathcal{F}_n = \mathbb{C}_{F_n}(x^\otimes)$  when  $x^\otimes = (0, 0, \dots, 0)$ .

Case 2: Suppose  $x^\otimes = (1, 1, \dots, 1)$ . Since  $\mathbb{C}_{F_n}(x^\otimes) = \{f^\wedge\}$  by (22), we need to show that  $\mathcal{F}_n^{x^\otimes} = \{f^\wedge\}$ . Initially, at time step 1,

$$F^{\downarrow x^\otimes}(1) = (f_1, z_1) = (f_1, x^\otimes) = ((0, 0, \dots, 0, 1), (1, 1, \dots, 1, 1))$$

then  $f_{i,j+k} > 0 \implies j+k = n$ ,  $\max\{\ell : z_{1,\ell} = 1\} = \max\{1, 2, \dots, n-1\} = n-1$ ,  $\hat{j} = \max\{\min\{n-1, n-1\}, \lceil \frac{n}{2} \rceil\} = n-1$  and

$$\Xi(f_1, z_1) = \{(j, k) : f_{i,j+k} > 0, 1 \leq j \leq n-1 \leq k \leq n-1\} = \{(1, n-1)\}.$$

Thus, the only state that is immediately preceded by  $(f_1, z_1)$  is our next state  $(f_2, z_2) = (f_1 - e_n + e_1 + e_{n-1}, z_1 - \mathbf{1}_{\{1\}}(z_{1,1})e_1 - \mathbf{1}_{\{1\}}(z_{i,n-1})e_{n-1})$  with probability 1 due to the equality of the numerator and denominator in (25):

$$(f_1, z_1) \prec_{f,z} (f_2, z_2) = ((1, 0, \dots, 1, 0), (0, 1, \dots, 1, 0)) = F^{\downarrow x^\otimes}(2)$$

In general, at time step  $i$ ,  $\Xi(f_i, z_i) = \{(1, n-i)\}$ ,  $P((f_{i+1}, z_{i+1}) | (f_i, z_i)) = 1$  and

$$f_{i+1} = f_1 - \sum_{j=1}^n e_j + \sum_{j=1}^i e_1 + \sum_{j=1}^i e_{n-j} = e_{n-i} + ie_1, \quad z_{i+1} = x^\otimes - e_1 - \sum_{j=i}^n e_{j-1}.$$

By (19),  $f_{i+1} = e_{n-i} + ie_1 = f_{i+1}^\wedge$  and we get the desired  $f$ -sequence  $f^\wedge = (f_n^\wedge, f_{n-1}^\wedge, \dots, f_1^\wedge)$  in the forward direction as the only realization over  $\mathbb{F}_n$  of our fully controlled Markov chain  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$ . Therefore,  $\mathcal{F}_n^{x^\otimes} = \{f^\wedge\} = \mathbb{C}_{F_n}(x^\otimes)$  when  $x^\otimes = (1, 1, \dots, 1)$ .

Case 3: Now, suppose  $x^\otimes \in \{0, 1\}^{n-1}$ . First, we will show that  $f \in \mathcal{F}_n^{x^\otimes}$  implies that  $f \in \mathcal{C}F_n(x^\otimes)$  or equivalently that  $f \notin F_n(x^\otimes)$ . We will prove by contradiction. Assume  $f \in \mathcal{F}_n^{x^\otimes}$ . Suppose that  $f \in F_n(x^\otimes)$ . Then by (21), there exists an  $h$  with  $x_h^\otimes = 1$  such that  $\sum_{i=1}^n f_{i,h} = 0$ . Since  $\sum_{i=1}^n f_{1,h} > 0$  and  $\sum_{i=1}^n f_{2,h} > 0$  for every  $f \in \mathcal{F}_n$ , with  $n > 2$ ,  $h \in \{3, 4, \dots, n-1\}$ . Recall that  $\sum_{i=1}^n f_{i,h} = 0$  implies that there was never a split of any lineage that birthed a child lineage subtending  $h$  leaves at any time step in the sequential realization of  $f = (f_1, f_2, \dots, f_n)$  over  $\mathbb{F}_n$  by  $\{F^{\downarrow x^\otimes}(k)\}_{[n]_+}$ . This contradicts our assumption that  $f \in \mathcal{F}_n^{x^\otimes}$  as it violates the constrained splitting imposed by  $\Xi(f_i, z_i)$  at the time step  $i$  when  $\max\{\ell : z_{i,\ell} = 1\} = h$  in the definition of  $\hat{j}$ . So, our supposition that  $f \in F_n(x^\otimes)$  is false. Therefore,  $f \in \mathcal{F}_n^{x^\otimes} \implies f \in \mathcal{C}F_n(x^\otimes)$ . Next, we will show  $f \in \mathcal{C}F_n(x^\otimes)$  implies that  $f \in \mathcal{F}_n^{x^\otimes}$ . Assume that  $f \in \mathcal{C}F_n(x^\otimes)$ , then  $\sum_{i=1}^n f_{i,h} > 0$  for every  $h \in \{h : x_h^\otimes = 1\}$  by (21). This means that for each  $h$  with  $x_h^\otimes = 1$  there is at least one split in  $f$  that birthed a child lineage subtending  $h$  leaves. Since this splitting condition satisfies the constraints imposed by  $\Xi(f_i, z_i)$  at each time step  $i$  when  $\max\{\ell : z_{i,\ell} = 1\} = h$ ,  $h \in \{h : x_h^\otimes = 1\}$ , in the definition of  $\hat{j}$ , this  $f$  can be sequentially realized over  $\mathbb{F}_n$  by  $\{F^{\downarrow x^\otimes}(k)\}_{[n]_+}$ . Therefore,  $f \in \mathcal{C}F_n(x^\otimes) \implies f \in \mathcal{F}_n^{x^\otimes}$ .

Thus, given  $\phi_1$  and an  $x^\otimes$ , we can efficiently propose SFS samples from  $\mathcal{X}_n^m$ , such that the underlying  $f$ -sequence  $f \in \mathcal{C}F_n(x^\otimes)$ , using Algorithm 2. Note however that a further straightforward importance sampling step using (25) and (11) is needed to obtain SFS samples that are distributed over  $\mathcal{X}_n^m$  according to the unlabeled  $n$ -coalescent over  $\mathcal{C}F_n(x^\otimes)$ . The number of sites  $m$  do not appear in either of the Algorithms in this article since  $\phi_1$  is the mutation rate of the entire locus.

---

**Algorithm 2** SFS Proposal under an  $x^\otimes$ -controlled unlabeled  $n$ -coalescent

---

1: **input:**

1. scaled mutation rate  $\phi_1$  of the locus
2. observed  $x^\otimes$  (note that sample size  $n = |x^\otimes| + 1$ )

2: **output:** an SFS sample  $x$  such that the underlying  $f$ -sequence  $f \in \mathcal{C}F_n(x^\otimes)$

3: generate an  $f$ -sequence  $f$  under  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$

4: draw  $t \sim T = (T_2, T_3, \dots, T_n) \sim \bigotimes_{i=2}^n \binom{i}{2} e^{-\binom{i}{2} t_i}$ , or as desired

5:  $l = t \cdot \mathbf{f}$ , where  $\mathbf{f} = \mathbf{F}(f)$

6: draw  $x$  from Poisson-Multinomial distribution  $e^{-\phi_1 l \cdot} (\phi_1 l \cdot)^s \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!$

7: **return:**  $x$

---

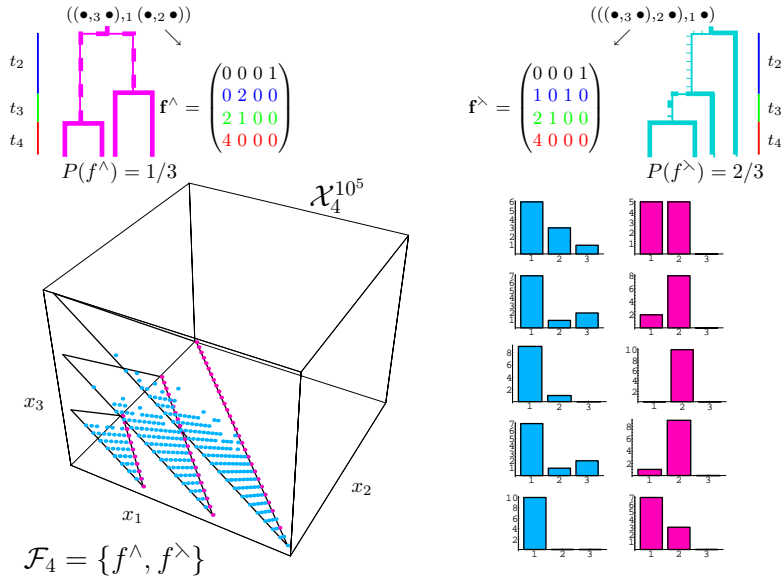
## 4 Applications

We briefly show how to apply Kingman’s unlabeled  $n$ -coalescent experiment and the associated Markov processes of §3 to estimate parameters and obtain the null distribution of test statistics based on simulated SFS data. A beta version of `LCE-0.1: A C++ class library for lumped coalescent experiments` that implements such algorithms is available from <http://www.math.canterbury.ac.nz/~r.sainudiin/codes/lce/> under the terms of the GNU General Public License.

### 4.1 Topologically-conditioned Tests

A large number of statistical tests on population-genetic data focus on summary statistics in lieu of the full data matrix, and estimate a (one- or two-tailed)  $p$ -value for that statistic under a model of interest. In the case of Tajima’s  $D$ , a statistic of the SFS [12, §3.3], simulations may be used to calculate  $Pr(D \leq d_{obs})$ , where  $d_{obs}$  is the observed value of  $D$  for a particular locus. The simulation procedure involves two steps. First, coalescent trees in  $\mathcal{C}_n \mathbb{T}_n$  are drawn randomly from the null model, with no respect to topological information contained in the full data matrix. Further, the observed number of mutations are placed onto each realized coalescent tree  $c^t$  [33]. In the empirical literature, there are a number of publications applying this procedure in order to identify “unusual” loci (reviewed in [34]). Such genome scans may be improved greatly at little additional computational cost by conditioning on the partial topological information contained in  $X^\circledast(x) = x^\circledast$  corresponding to the SFS  $x$  at a locus that is assumed to be free of intra-locus recombination and evolving neutrally under the standard null hypothesis. Using Algorithm 2 we can obtain topologically-conditioned null distributions of test statistics that are functions of the SFS.

Figure 6 illustrates the problem of ignoring the topological information in  $x^\circledast$ , when it is readily available, even when  $n = 4$ . Notice that 12 out of the 18  $c$ -sequences in  $\mathcal{C}_4$  have unbalanced trees that map to  $f^\lambda$  and the remaining 6  $c$ -sequences have balanced trees that map to  $f^\wedge$ . Recall that Kingman’s labeled  $n$ -coalescent assigns the uniform distribution over  $\mathcal{C}_n$ , while  $P(f)$  for any  $f \in \mathcal{F}_n$  is far from uniformly distributed under the Kingman’s unlabeled  $n$ -coalescent and easily obtained from (8) or (11). We can also obtain  $P(f)$  from higher-order shape statistics of the  $f$ -sequence [28, Prop. 3.31]. Thus,  $P(c) = 1/18$  for each  $c \in \mathcal{C}_n$  while  $P(f^\lambda) = 2/3$  and  $P(f^\wedge) = 1/3$ . Five SFS simulations upon  $f^\lambda$  and  $f^\wedge$  are shown as the left and right columns of bar charts, respectively, on the lower right corner of Figure 6. The remaining simulated SFS are plotted in the simplex with a fixed number of segregating sites  $s = \sum_{i=1}^{n-1} x_i$  contained in  $\mathcal{X}_4^{10^5}$ , the sample space of SFS with four sampled individuals at  $10^5$  sites. Observe how every SFS simulated under  $f^\wedge$  has  $x_3 = 0 \implies x_3^\circledast = 0$ , as opposed to those SFS simulated under  $f^\lambda$ . Crucially, if we do not know the hidden  $f \in \{f^\lambda, f^\wedge\}$  that the observed SFS  $x$  was realized upon, then the observation that  $x_3 > 0 \implies x_3^\circledast = 1$ , and this allows us to unambiguously eliminate  $f^\wedge$  from the hidden space of  $f$ -sequences we need to integrate over or



**Fig. 6.** Topological Unfolding of SFS and Tajima's  $D$  when  $n = 4$

conditionally simulate from. This set of  $x^\otimes$ -specific hidden  $f$ -sequences is exactly  $\mathcal{C}_{F_n}(x^\otimes)$  that we can access with the proposal Markov chain  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$  and its importance-reweighed variants. Thus, by means of Algorithm 2 that invokes  $\{F^{\downarrow x^\otimes}(k)\}_{k \in [n]_+}$  and further reweighing by  $P(f)$  we can generate the null distribution of any statistic that is a function of SFS. Such SFS statistics include the classical linear combinations covered in the companion article [12] as well as various classical and non-classical tree shape statistics [28, §4.3].

## 4.2 Parameter Estimation in an Exponentially Growing Population

We estimate the scaled mutation rate  $\phi_1^*$  and the exponential growth rate  $\phi_2^*$  based on the observed SFS at one non-recombining locus of length  $m$  from  $n$  samples. The performance of our estimator is assessed over 1,000 data sets that were simulated under our model with the locus-specific scaled mutation rate  $\phi_1^* = 10.0$ , constant population size with the growth rate  $\phi_2^* = 0.0$  (for human data  $\phi_1^* = 10.0$  implies a locus of length 100kbp, i.e.  $m = 100,000$ ). Our choices of  $\phi_1^*$  and  $m$  are biologically motivated by a previous study on human SNP density [35]. We choose  $\phi_2^* = 0.0$  for the standard  $n$ -coalescent null model. Our point estimate  $(\widehat{\phi}_1, \widehat{\phi}_2)$  of  $(\phi_1^*, \phi_2^*)$  based on the SFS  $x$  is the maximum *a posteriori* estimate obtained from a histogram estimate of the posterior  $P(\phi|x)$ . The histogram is based on a uniform grid of  $101 \times 101$  parameter points  $\phi = (\phi_1, \phi_2)$  over our rectangular uniform prior density  $((100 - 1/10000)100)^{-1} \mathbf{1}_{\{[0.0001, 100], [0, 100]\}}(\phi_1, \phi_2)$ .

There are several quantities one can use to gauge the efficiency of our estimator  $\widehat{\phi}$  at this point. Our performance measures based on 1000 replicates of

$n$	$\widehat{\phi}_2$			$\widehat{\phi}_1$			$(\widehat{\phi}_1, \widehat{\phi}_2)$	
	$\sqrt{se}$	$bs$	$C_{99\%}$	$\sqrt{se}$	$bs$	$C_{99\%}$	$C_{99\%}$	$Qrt(\check{\mathcal{K}})$
4	46	30	42	43	30	53	98	{0.061, 0.079, 0.13}
5	32	19	42	31	22	63	96	{0.074, 0.098, 0.16}
6	31	18	41	35	23	69	93	{0.082, 0.11, 0.17}
7	34	19	48	32	20	68	87	{0.090, 0.12, 0.21}
8	26	12	66	21	11	72	92	{0.098, 0.14, 0.26}
9	27	12	65	18	10	70	93	{0.097, 0.14, 0.21}
10	23	11	64	17	10	66	95	{0.091, 0.14, 0.30}

**Table 1.** Performance of our estimator of  $\phi_1^*$  and  $\phi_2^*$  based on SFS (see text).

SFS simulated under  $\phi_1^* = 10.0$  and  $\phi_2^* = 0.0$  can help make natural connections to the theory of approximate sufficiency [36], as we not only measure the bias ( $bs$ ), root-mean-squared-error ( $\sqrt{se}$ ) and the marginal and joint 99% empirical coverage ( $C_{99\%}$ ) but also the data-specific variation in the concentration of the posterior distribution as summarized by the quartiles of  $\check{\mathcal{K}}$ , the Kullback-Leibler divergence between the posterior histogram estimate and the uniform prior that is rescaled by the prior’s entropy. Table 1 gives the maximum *a posteriori* estimates of  $\phi = (\phi_1, \phi_2)$  by a Monte Carlo sum over  $\phi$ -specific  $t$ ’s in  $\mathbb{T}_n := (0, \infty)^{n-1}$  that estimates (24).

#### ACKNOWLEDGMENTS

R.S. was supported by a research fellowship from the Royal Commission for the Exhibition of 1851 under the sponsorship of Peter Donnelly during the course of this study. R.S. thanks Mike Steel for [3, def. 6.3.1], Jesse Taylor for [1, 5.2], Joe Watkins for the articulation of def. 2, Michael Nussbaum and Simon Tavaré for discussions on approximate sufficiency and Scott Williamson for generous introductions to the site frequency spectrum.

#### References

1. Kingman, J.: On the genealogy of large populations. *Journal of Applied Probability* **19** (1982) 27–43
2. Kingman, J.: The coalescent. *Stochastic Processes and their Applications* **13** (1982) 235–248
3. Kemeny, Snell: *Finite Markov Chains*. D. Van Nostrand Co. (1960)
4. Griffiths, R., Tavaré, S.: The genealogy of a neutral mutation. In Green, P., Hjort, N., Richardson, S., eds.: *Highly Structured Stochastic Systems*. Oxford University Press (2003) 393–412
5. Yang, Z.: Complexity of the simplest phylogenetic estimation problem. *Proceedings Royal Soc. London B Biol. Sci.* **267** (2000) 109–119
6. Hosten, S., Khetan, A., Sturmfels, B.: Solving the likelihood equations. *Found. Comput. Math.* **5** (2005) 389–407
7. Casanellas, M., Garcia, L., Sullivant, S.: Catalog of small trees. In Pachter, L., Sturmfels, B., eds.: *Algebraic statistics for computational biology*. Cambridge University Press (2005) 291–304

8. Sainudiin, R., York, T.: Auto-validating von neumann rejection sampling from small phylogenetic tree spaces. *Algorithms for Molecular Biology* **4** (2009) 1
9. Weiss, G., von Haeseler, A.: Inference of population history using a likelihood approach. *Genetics* **149** (1998) 1539–1546
10. Beaumont, M., Zhang, W., Balding, D.: Approximate Bayesian computation in population genetics. *Genetics* **162** (2002) 2025–2035
11. Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S.: Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** (2003) 15324–15328
12. Sainudiin, R., Thornton, K., Booth, J., Stillman, M., Yoshida, R.: Coalescent experiments II: Markov bases of classical population genetic statistics – UCDMS Research Report 2009/8, may 19, 2009 (submitted). Available at <http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/CoalExpsII.pdf> (2009)
13. Watterson, G.: On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7** (1975) 256–276
14. Griffiths, R., Tavaré, S.: Ancestral inference in population genetics. *Stat. Sci.* **9** (1994) 307–319
15. Griffiths, R., Tavaré, S.: Markov chain inference methods in population genetics. *Math. Comput. Modelling* **23** (1996) 141–158
16. Bahlo, M., Griffiths, R.: Inference from gene trees in a subdivided population. *Theoret. Pop. Biol.* **57** (1996) 79–95
17. Stephens, M., Donnelly, P.: Inference in molecular population genetics. *J. R. Statist. Soc. B* **62** (2000) 605–655
18. Slatkin, M.: A vectorized method of importance sampling with applications to models of mutation and migration. *Theoret. Pop. Biol.* **62** (2002) 339–348
19. Iorio, M., Griffiths, R.: Importance sampling on coalescent histories. I. *Adv. Appl. Prob.* **36** (2004) 417–433
20. Kolmogorov, A.: Sur l'estimation statistique des paramètres de la loi de gauss. *Bull. Acad. Sci. URSS Ser. Math.* **6** (1942) 3–32
21. Felsenstein, J.: Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution* **23** (2006) 691700
22. Tavaré, S.: Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology* **26** (1984) 119–164
23. Kendall, D.: Some problems in mathematical genealogy. In Gani, J., ed.: *Perspectives in Probability and Statistics*. Academic Press (1975) 325–345
24. Fisher, R.: *The Genetical Theory of Natural Selection*. Clarendon, Oxford (1930)
25. Wright, S.: Evolution in mendelian populations. *Genetics* **16** (1931) 97–159
26. Ewens, W.: The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* **3** (1972) 87–112
27. Ewens, W.: A note on the sampling theory of infinite alleles and infinite sites models. *Theoret. Pop. Biol.* **6** (1974) 143–148
28. Sainudiin, R., Stadler, T.: A unified multi-resolution coalescent: Markov lumpings of the Kingman-Tajima  $n$ -coalescent – UCDMS Research Report 2009/4, april 5, 2009 (submitted). Available at <http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/SixCoal.pdf> (2009)
29. Semple, C., Steel, M.: *Phylogenetics*. Oxford University Press (2003)
30. Rosenblatt, M.: *Random Processes*. Springer-Verlag (1974)
31. Siek, J.G., Lee, L.Q., Lumsdaine, A.: *The Boost Graph Library User Guide and Reference Manual (With CD-ROM)*. Addison-Wesley Professional (2001)
32. Dufflo, M.: *Random Iterative Models*. Springer (1997)
33. Hudson, R.: The how and why of generating gene genealogies. In Clark, A., Takahata, N., eds.: *Mechanisms of Molecular Evolution*. Sinauer (1993) 23–36

34. Thornton, K., Jensen, J.D., Becquet, C., Andolfatto, P.: Progress and prospects in mapping recent selection in the genome. *Heredity* **98** (2007) 340–348
35. Sainudiin, R., Clark, A., Durrett, R.: Simple models of genomic variation in human SNP density. *BMC Genomics* **8** (2007) 146
36. Cam, L.L.: Sufficiency and approximate sufficiency. *Ann. Math. Stats.* **35** (1964) 1419–1455