Bull Math Biol (2011) 73: 829–872
DOI 10.1007/s11538-010-9605-5

ORIGINAL ARTICLE

Experiments with the Site Frequency Spectrum

Raazesh Sainudiin · Kevin Thornton · Jennifer Harlow · James Booth · Michael Stillman · Ruriko Yoshida · Robert Griffiths · Gil McVean · Peter Donnelly

Received: 15 May 2009 / Accepted: 2 November 2010 / Published online: 23 December 2010

Abstract Evaluating the likelihood function of parameters in highly-structured population genetic models from extant deoxyribonucleic acid (DNA) sequences is computationally prohibitive. In such cases, one may approximately infer the parameters from summary statistics of the data such as the site-frequency-spectrum (SFS) or its linear combinations. Such methods are known as approximate likelihood or Bayesian computations. Using a controlled lumped Markov chain and computational commutative algebraic methods, we compute the exact likelihood of the SFS and many classical linear combinations of it at a non-recombining locus that is neutrally evolving

R. Sainudiin (✉)
Biomathematics Research Centre, Private Bag 4800, Christchurch 8041, New Zealand
e-mail: r.sainudiin@math.canterbury.ac.nz

Present address:
R. Sainudiin
Chennai Mathematical Institute, Plot H1, SIPCOT IT Park, Padur PO, Siruseri 603103, India

K. Thornton
Department of Ecology and Evolutionary Biology, University of California, Irvine, USA

J. Harlow · R. Sainudiin
Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

J. Booth
Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, USA

M. Stillman
Department of Mathematics, Cornell University, Ithaca, USA

R. Yoshida
Department of Statistics, University of Kentucky, Lexington, USA

R. Griffiths · G. McVean · P. Donnelly
Department of Statistics, University of Oxford, Oxford, UK

under the infinitely-many-sites mutation model. Using a partially ordered graph of coalescent experiments around the SFS, we provide a decision-theoretic framework for approximate sufficiency. We also extend a family of classical hypothesis tests of standard neutrality at a non-recombining locus based on the SFS to a more powerful version that conditions on the topological information provided by the SFS.

Keywords Controlled lumped coalescent · Population genetic Markov bases

1 Introduction

Models in population genetics are highly structured stochastic processes (Griffiths and Tavare 2003). Inference is typically conducted with data that is modeled as a partial observation of one realization of such a process. Likelihood methods are most 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 *multiple sequence alignment* or *MSA* data $u_o \in \mathcal{U}_n^m$ that was observed at the finest available empirical resolution, given a parameter $\phi \in \boldsymbol{\Phi}$, that is indexing a biologically motivated family of models. The MSA sample space $\mathcal{U}_n^m := \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^{n \times m}$ 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 (Yang 2000; Hosten et al. 2005; Casanellas et al. 2005; Sainudiin and York 2009). Computationally-intensive inference, based on an observed $u_o \in \mathcal{U}_n^m$, with realistically large n and m, is currently impossible for recombining loci and prohibitive for non-recombining loci.

An alternative inference strategy that is computationally feasible involves a relatively low-dimensional statistic $R(u_o) = r_o \in \mathcal{R}_n^m$ of $u_o \in \mathcal{U}_n^m$. 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 of the observed data u_o. Since R is typically not a sufficient statistic for ϕ, i.e., $P(\phi|r) \neq P(\phi|u)$. Such methods have been termed as *approximate likelihood computations or ALC* (Weiss and von Haeseler 1998) in a frequentist setting and as *approximate Bayesian computations or ABC* (Beaumont et al. 2002) in a Bayesian setting. ALC and ABC are popular simulation-based inference methods in computational population genetics as they both provide an easily implementable inference procedure for any model that you can simulate from. Several low dimensional (summary) statistics, each of which are not shown to be sufficient or even necessarily consistent, form the basis of information in such approximate likelihood or Bayesian computations. The underlying assumption that ensures asymptotic consistency of this estimator is that a large enough set of such statistics will be a good

proxy for the observed data u_o, in an approximately sufficient sense. However, there are several senses in which a set of population genetic statistics can be *large enough* for asymptotically consistent estimation. Furthermore, any formal notion of approximate sufficiency in population genetic experiments must account for the fact that the likelihood is defined by the n-coalescent prior mixture over elements in a partially observed genealogical space ${}^{\mathcal{C}_n}\mathbb{T}_n$:

$$P(r_o|\phi) = \int_{{}^c t \in {}^{\mathcal{C}_n}\mathbb{T}_n} P\left(r_o|{}^c t, \phi\right) dP\left({}^c t|\phi\right). \tag{1}$$

The discrete aspects of this hidden space account for the sequence of coalescence events, while the continuous aspects account for the number of generations between such events in units of rescaled time. We formalize at least three notions or senses of asymptotic consistency for various statistics of the data using a graph of partially ordered coalescent experiments under Watterson's infinitely-many-sites (IMS) model of mutation (Watterson 1975) and show that asymptotic consistency does not hold in every sense for the site frequency spectrum (SFS), a popular summary statistic of the MSA data, and its linear combinations, unless one can appropriately integrate over $\{{}^c t \in {}^{\mathcal{C}_n}\mathbb{T}_n : P(r_o|{}^c t, \phi) > 0\}$ in (1). This elementary observation has cautionary implications for simulation-intensive parameter estimation using ABC or ALC methods as well as outlier-detection using genome scanning methods that attempt to reject loci that are hypothesised to evolve under the standard neutral null model.

Our first specific objective here is to address the problem of inferring the posterior distribution over the same parameter space $\boldsymbol{\Phi}$ across different empirical resolutions or statistics of n DNA sequences with m homologous sites drawn from a large Wright–Fisher population at a large non-recombining locus that is neutrally evolving under the infinitely-many-sites model of mutation. The empirical resolutions of interest at the coarsest end, include classical statistics, such as (i) the nonnegative integer-valued *number of segregating sites* $S \in \mathbb{Z}_+$ (Watterson 1975), (ii) the rational-valued *average heterozygosity* $\pi \in \mathbb{Q}$, (iii) the real-valued Tajima's D (Tajima 1989) that combines (i) and (ii). At a slightly finer resolution than the first three that is of interest is (iv) the nonnegative integer vector called the *folded site frequency spectrum* $y \in \mathbb{Z}_+^{\lfloor n/2 \rfloor}$. At an intermediate resolution, (v) the nonnegative integer vector called the *site frequency spectrum* $x \in {\mathbb{Z}_+}^{n-1}$ is a much finer statistic whose linear combinations determine (i), (ii), (iii), and (iv), in addition to various other statistics in the literature, including folded singletons $y_1 := x_1 + x_{(n-1)}$ (Hudson 1993) and Fay and Wu's $\theta_H := (n(n-1))^{-1} \sum_{i=1}^{n-1} (2\, i^2\, x_i)$ (Fay and Wu 2000). See Wakeley (2007) for a discussion of the linear relations between various classical summaries and the site frequency spectrum. At the finest resolution we can conduct inference on the basis of (vi) binary incidence matrices that are sufficient for the infinitely-many-sites model of mutation using existing methods (for, e.g., Stephens and Donnelly 2000). The asymptotic consistency emphasised here involves a single locus, that is free of intralocus recombination across n individuals and at m homologous sites, as m approaches infinity.

Our second specific objective here is to extend a class of hypothesis tests of the standard neutral model for a non-recombining locus toward the intermediate empirical resolution of the SFS. This class includes various classical "Tajima-like" tests in

the sense of Ewens (2000, p. 361) as well as others that are based on the null distribution of the SFS. Our extension involves conditioning the null distribution by an equivalence class of unlabeled coalescent tree topologies, up to a partial information provided by the observed SFS. Thus, the null distribution over the SFS sample space, that in turn determines the null distributions of all the test statistics in our class, are only based on those genealogies whose coalescent tree topologies have a nonzero probability of underlying our observed SFS. This amounts to an "unlabeled topological conditioning" of any test statistic for neutrality that is a function of the site frequency spectra, including several classical tests.

Two elementary ideas form the basic structures that are exploited in this paper to achieve the objectives outlined in the previous two paragraphs. First, we develop a Markov lumping of Kingman's n-coalescent to Kingman's *unlabeled n-coalescent* as suggested in Kingman (1982b, (5.1), (5.2)) but without explicit pursuit. The unlabeled n-coalescent is a Markov chain on a many-to-one map of the state space of the n-coalescent (or more specifically, the labeled n-coalescent) and it is sufficient and necessary to prescribe the Φ-indexed family of measures for the sample space of the SFS. Secondly, we exactly evaluate the posterior density based on one or more linear combinations of the observed site frequency spectrum. This is accomplished by an elementary study of the algebraic geometry of such statistics using Markov bases (Diaconis and Sturmfels 1998). A beta version of `LCE-0.1: A C++ class library for lumped coalescent experiments` that implements such algorithms is publicly available from http://www.math.canterbury.ac.nz/~r.sainudiin/codes/lce/ under the terms of the GNU General Public License.

2 Genealogical and Mutational Models

The stochastic models for the genealogy of the sample and the mutational models that generate data are given in this section.

2.1 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 $\phi_2 = 0$, each offspring "chooses" its parent 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:

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

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 ancestral lineages in the previous generation is:

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

Let $\mathbb{Z}_- := \{0, -1, -2, \ldots\}$ 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 $\mathbb{H}_n := \{n, n-1, \ldots, 1\}$ with 1-step transition probabilities given by (2). $\{{}^{N}H^{\uparrow}(k)\}_{k\in\mathbb{Z}_-}$ is 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: $(1 - 1/N)^{\lfloor Nt \rfloor} \overset{N\to\infty}{\longrightarrow} e^{-t}$.

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 in (2), as the population size N tends to infinity:

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

$q_{i,j} = 0$ 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 ith 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 $T = (T_2, T_3, \ldots, T_n)$ 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 $\prod_{i=2}^{n}\binom{i}{2}e^{-\binom{i}{2}t_i}$ 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, $[n]_- := \{n, n-1, \ldots, 2, 1\}$ denotes the decreasingly ordered discrete time index set. Similarly, let $[n]_+ := \{1, 2, \ldots, n-1, n\}$ denote the increasingly ordered discrete time index set.

2.2 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 partitions of the label set $\mathfrak{L} = \{1, 2, \ldots, n\}$ of our n samples. Denote by $\mathbb{C}_n^{(i)}$ the set of all partitions with i blocks, i.e., $\mathbb{C}_n = \bigcup_{i=1}^{n} \mathbb{C}_n^{(i)}$. Let $c_i := \{c_{i,1}, c_{i,2}, \ldots, c_{i,i}\} \in \mathbb{C}_n^{(i)}$ denote the i elements of c_i. The *labeled n-coalescent partial ordering* on $\mathbb{C}_n$ is based on the immediate precedence relation $\prec_c$:

$$c_{i'} \prec_c c_i \quad \Longleftrightarrow \quad c_{i'} = c_i \setminus c_{i,j} \setminus c_{i,k} \cup (c_{i,j} \cup c_{i,k}), \quad j \neq k, j, k \in \left\{1, 2, \ldots, |c_i|\right\}.$$

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 $C^{\uparrow}(n) = c_n = \{\{1\}, \{2\}, \ldots, \{n\}\}$ and final absorbing state $C^{\uparrow}(1) = c_1 = \{\{1, 2, \ldots, n\}\}$, with the following transition probabilities Kingman (1982a, Eq. (2.2)):

$$P(c_{i'}|c_i) = \begin{cases} \binom{i}{2}^{-1} & \text{:if } c_{i'} \prec_c c_i,\ c_i \in \mathbb{C}_n^{(i)}, \\ 0 & \text{:otherwise.} \end{cases} \tag{3}$$

Now, 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 realization of the chain, and denote the space of such c-sequences by

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

The probability that $c_i \in \mathbb{C}_n^{(i)}$ is visited by the chain Kingman (1982a, Eq. (2.3)) is:

$$P(c_i) = \frac{(n-i)!i!(i-1)!}{n!(n-1)!} \prod_{j=1}^{i} |c_{i,j}|, \tag{4}$$

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

$$P(c) = \prod_{i=n}^{2} P(c_{i-1}|c_i) = \frac{2^{n-1}}{n!(n-1)!} = \frac{1}{|\mathcal{C}_n|}. \tag{5}$$

Kingman's labeled n-coalescent (Kingman 1982a, 1982b) denoted by $\{C^{\uparrow}(t)\}_{t \in \mathbb{R}_+}$, is a continuous-time Markov chain 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 (Kingman 1982b, Eq. (2.10)):

$$q(c_{i'}|c_i) = \begin{cases} -\binom{i}{2} & \text{:if } c_i = c_{i'},\ c_i \in \mathbb{C}_n^{(i)}, \\ 1 & \text{:if } c_{i'} \prec_c c_i, \\ 0 & \text{:otherwise.} \end{cases} \tag{6}$$

The above instantaneous transition rates for $\{C^{\uparrow}(t)\}_{t\in\mathbb{R}_+}$ are obtained by an independent coupling of the death process $\{H^{\uparrow}(t)\}_{t\in\mathbb{R}_+}$ in Sect. 2.1 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 $\mathfrak{L}$ and taken at random from the Wright–Fisher population of constant size N. This asymptotic approximation, as the population size $N \to \infty$, can be seen using arguments similar to those in Sect. 2.1. See Kingman (1982a, Sects. 1–2) for this construction.

Let the space of *ranked, rooted, binary, phylogenetic trees* with leaves or samples labeled by $\mathfrak{L} = \{1, 2, \ldots, n\}$ (Semple and Steel 2003, §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

$$^{\mathcal{C}_n}\mathbb{T}_n := \mathcal{C}_n \otimes \mathbb{T}_n := \left\{ {}^{c}t := \left({}^{c_n}t_n, {}^{c_{n-1}}t_{n-1}, \ldots, {}^{c_2}t_2 \right) : c \in \mathcal{C}_n, t \in \mathbb{T}_n := \mathbb{R}_+^{n-1} \right\}.$$

$^{\mathcal{C}_n}\mathbb{T}_n$ 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. Figure 1 depicts the n-coalescent tree space $^{\mathcal{C}_3}\mathbb{T}_3$ for the sample label set $\mathfrak{L} = \{1, 2, 3\}$ with sample size $n = 3$.

2.3 Kingman's Unlabeled n-Coalescent

Next, we model the sample genealogy at a resolution that is finer than the number of ancestral lineages but coarser than that of the labeled n-coalescent. This is 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-destroyed" process by Kingman (1982b, 5.2). Tavaré (1984, pp. 136–137) terms it the "family-size process" along the nomenclature of a more general birth-death-immigration process (Kendall 1975). The transition probabilities of this Markov process, in either temporal direction, are not explicitly developed in Kingman (1982b) or Tavaré (1984). 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 $f_i := (f_{i,1}, f_{i,2}, \ldots, f_{i,n})$. Then the space of f_i's is defined by

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

Let the set of such frequencies over all epochs be $\mathbb{F}_n := \bigcup_{i=1}^{n} \mathbb{F}_n^{(i)}$. Let us define an *$f$-sequence* f as

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

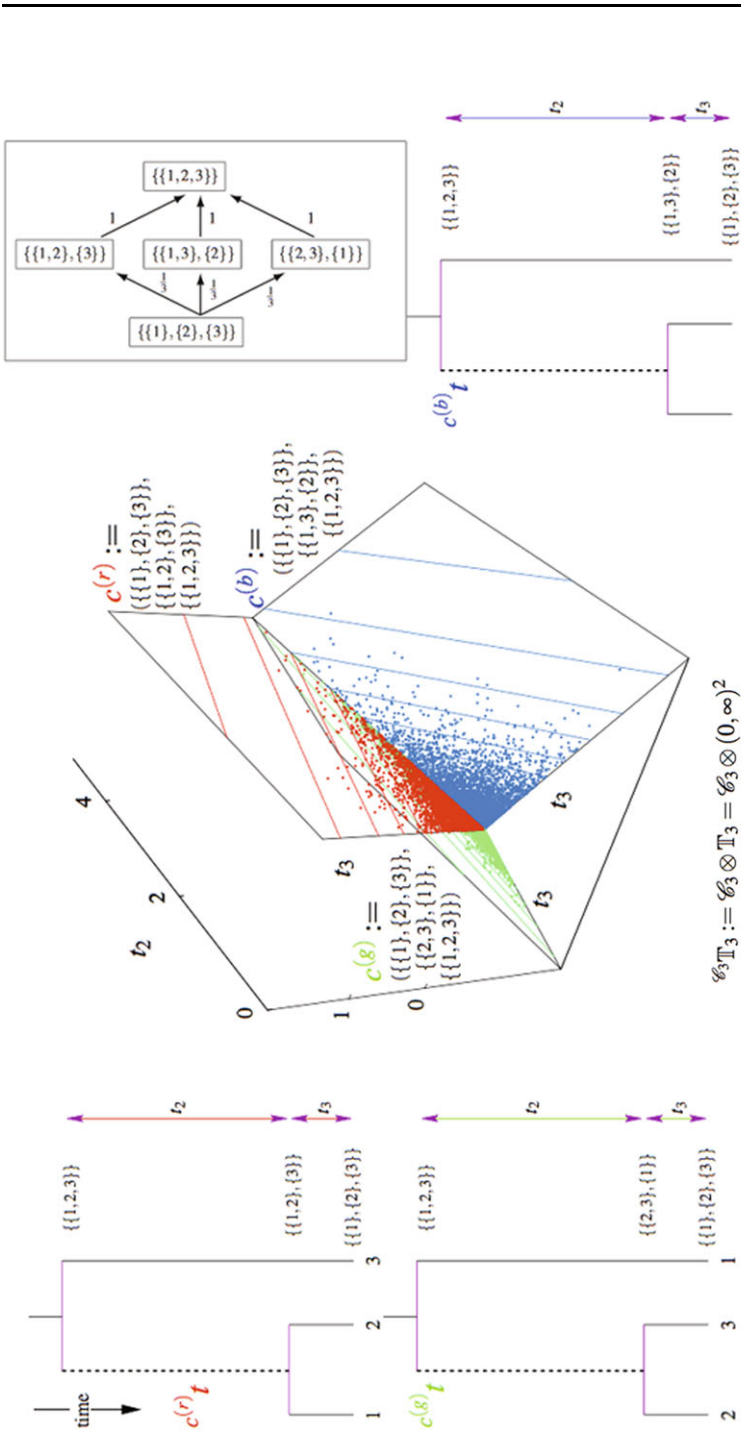

Fig. 1 Realizations of 3-coalescent trees in the space of such trees is plotted on the three rectangles as colored points in middle panel. The lines on the rectangles are the contours of the independent exponentially distributed epoch times for each c-sequence. Each of the three coalescent trees, with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

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

$$f_{i'} \prec_f f_i \quad \Longleftrightarrow \quad f_{i'} = f_i - e_j - e_k + e_{j+k}. \tag{7}$$

Thus, $\mathcal{F}_n$ is the space of f-sequences with n samples, i.e., the space of the frequencies of the cardinalities of c-sequences in $\mathcal{C}_n$. Recall the c-sequence $c = (c_n, c_{n-1}, \ldots, c_1)$, 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 $c_i := \{c_{i,1}, c_{i,2}, \ldots, c_{i,i}\}$ contains i subsets. 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 $\mathbb{1}_A(a) = 0$). Then the corresponding f-sequence is given by the map $\underline{\mathscr{F}}(c) = f : \mathcal{C}_n \to \mathcal{F}_n$, as follows:

$$\begin{aligned} \underline{\mathscr{F}}(c) &:= \big(\mathscr{F}(c_n), \ldots, \mathscr{F}(c_1)\big), \\ \mathscr{F}(c_i) &:= \left(\sum_{h=1}^{i} \mathbb{1}_{\{1\}}\big(|c_{i,h}|\big), \ldots, \sum_{h=1}^{i} \mathbb{1}_{\{n\}}\big(|c_{i,h}|\big) \right). \end{aligned} \tag{8}$$

Thus, $\mathcal{F}_n$ indexes an equivalence class in $\mathcal{C}_n$ via $\underline{\mathscr{F}}^{[-1]}(f)$, the inverse map of (8). 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 $\{C^{\uparrow}(k)\}_{k\in[n]_-}$ on $\mathbb{C}_n$ given by (3). $\{F^{\uparrow}(k)\}_{k\in[n]_-}$ is the embedded discrete time Markov chain of the unlabeled n-coalescent.

Proposition 1 (Backward Transition Probabilities of an f-sequence) *The probability of $f := (f_n, f_{n-1}, \ldots, f_1) \in \mathcal{F}_n$ under the n-coalescent is given by the product*:

$$P(f) = \prod_{i=n}^{2} P(f_{i-1}|f_i), \tag{9}$$

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 $f_i \in \mathbb{F}_n^{(i)}$, $f_{i-1} \in \mathbb{F}_n^{(i-1)}$:

$$P(f_{i-1}|f_i) = \begin{cases} f_{i,j} f_{i,k} \binom{i}{2}^{-1} & :\text{if } f_{i-1} = f_i - e_j - e_k + e_{j+k},\ j \neq k, \\ \binom{f_{i,j}}{2} \binom{i}{2}^{-1} & :\text{if } f_{i-1} = f_i - e_j - e_k + e_{j+k},\ j = k, \\ 0 & :\text{otherwise} \end{cases} \tag{10}$$

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

Proof Since (9) is obtained from (10) by Markov property, we prove (10) 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 $i \in \{n, n-1, \ldots, 3, 2\}$ ensures that our f remains in $\mathcal{F}_n$ as we go backwards in time from the nth coalescent epoch with n samples to the first one with the single ancestral lineage. □

The next proposition is a particular case of Tavaré (1984, Eq. (7.11)). We state and prove it here in our notation using coalescent arguments for completeness.

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 is*

$$P(f_i) = \frac{i!}{\prod_{j=1}^{i} f_{i,j}!} \binom{n-1}{i-1}^{-1}. \tag{11}$$

Proof Recall that $f_{i,j}$ is the number of edges that subtend j leaves during the ith coalescent epoch, where, $j \in \{1, 2, \ldots, n\}$. Now, label the i edges in some arbitrary manner. Let the number of the subtended leaves from the i labeled edges be $\Lambda := (\Lambda_1, \Lambda_2, \ldots, \Lambda_i)$. Due to the n-coalescent, Λ is a random variable with a uniform distribution on integer partitions of n, such that $\sum_{j=1}^{i} \Lambda_i = n$ and $\Lambda_i \geq 1$. 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* $f := (f_n, f_{n-1}, \ldots, f_1) \in \mathcal{F}_n$ *is given by the product*:

$$P(f) = \prod_{i=2}^{n} P(f_i | f_{i-i}), \tag{12}$$

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* $[n]_+ := \{1, 2, \ldots, n\}$:

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

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

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

Proof Since (12) follows from (13) due to Markov property, we prove (13) next. An application of the definition of conditional probability twice, followed by (2 yields):

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

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 $f_{i,h} = f_{i-1,h}, \forall h \in \{1, 2, \ldots, n\} \setminus \{j, k, j + k\}$.

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

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

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

This concludes the proof. □

Kingman's unlabeled n-coalescent or the *unvintaged and sized n-coalescent* in the descriptive nomenclature of Sainudiin and Stadler (2009) 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

$$q(f_{i'}|f_i) = \begin{cases} -i(i-1)/2 & :\text{if } \mathbb{F}_n^{(i)} \ni f_i = f_{i'}, \\ f_{i,j} f_{i,k} & :\text{if } \mathbb{F}_n^{(i-1)} \ni f_{i'} = f_i - e_j - e_k + e_{j+k}, \\ & j \neq k,\ f_i \in \mathbb{F}_n^{(i)}, \\ (f_{i,j})(f_{i,j}-1)/2 & :\text{if } \mathbb{F}_n^{(i-1)} \ni f_{i'} = f_i - e_j - e_k + e_{j+k}, \\ & j = k,\ f_i \in \mathbb{F}_n^{(i)}, \\ 0 & :\text{otherwise.} \end{cases} \tag{14}$$

The initial state is $f_n = (n, 0, 0, \ldots, 0)$ and the final absorbing state is $f_1 = (0, 0, \ldots, 1)$. 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 Sect. 2.1 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 Sainudiin and Stadler 2009, Proposition 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 (14), i.e.,*

$${}^N P_{f_i,f_{i'}}\left(\lfloor Nt \rfloor\right) \overset{N\to\infty}{\longrightarrow} P_{f_i,f_{i'}}(t) = \exp(Qt).$$

Proof For a proof see Sainudiin and Stadler (2009). □

Remark 1 (Markovian lumping from $\mathbb{C}_n$ to $\mathbb{F}_n$ via $\mathscr{F}$) 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 $\mathscr{F}$, is Markov as pointed out by Kingman (1982b, (5.1), (5.2)) using the arguments in Rosenblatt (1974, Sect. IIId). See Sainudiin and Stadler (2009) for an introduction to lumped coalescent processes and a proof that $\{F^{\uparrow}(t)\}_{t\in\mathbb{R}_+}$ is a Markov lumping of $\{C^{\uparrow}(t)\}_{t\in\mathbb{R}_+}$.

First, we introduce a matrix form $\mathbf{f}$ of f. Any f-sequence $f = (f_n, f_{n-1}, \ldots, f_1)$, 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:

$$\mathbf{F} : \mathcal{F}_n \to \mathbb{Z}_+^{(n-1)\times(n-1)}, \quad \mathbf{F}(f) = \mathbf{f} := \begin{pmatrix} f_{2,1} & f_{2,2} & \cdots & f_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n-1,1} & f_{n-1,2} & \cdots & f_{n-1,n-1} \\ f_{n,1} & f_{n,2} & \cdots & f_{n,n-1} \end{pmatrix}. \tag{15}$$

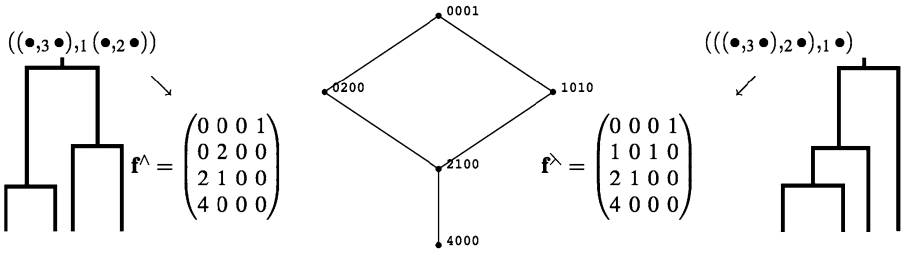

Fig. 2 The two f-sequences $f^{\rangle}$ and $f^{\wedge}$ corresponding to the balanced (*left panel*) and unbalanced unlabeled genealogies of four samples (*right panel*) are depicted as f-matrices $\mathbf{f}^{\wedge}$ and $\mathbf{f}^{\rangle}$, respectively. Hasse diagram of the state transition diagrams of $\{F^{\uparrow}(k)\}_{k\in[4]_-}$ and $\{F^{\downarrow}(k)\}_{k\in[4]_+}$ on $\mathbb{F}_4$ (*middle panel*)

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

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

Example 1 (Three Samples) When there are three samples we have three c-sequences: $c^{(r)}$, $c^{(b)}$ and $c^{(g)}$ (see Fig. 1) and all of them map to the only f-sequence f:

$$\begin{aligned} f &= \big((3,0,0),(1,1,0),(0,0,1)\big) \\ &= \underline{\mathscr{F}}\big(c^{(r)}\big) := \underline{\mathscr{F}}\Big(\big(\{\{1\},\{2\},\{3\}\},\{\{1,2\},\{3\}\},\{\{1,2,3\}\}\big)\Big) \\ &= \underline{\mathscr{F}}\big(c^{(b)}\big) := \underline{\mathscr{F}}\Big(\big(\{\{1\},\{2\},\{3\}\},\{\{1,3\},\{2\}\},\{\{1,2,3\}\}\big)\Big) \\ &= \underline{\mathscr{F}}\big(c^{(g)}\big) := \underline{\mathscr{F}}\Big(\big(\{\{1\},\{2\},\{3\}\},\{\{2,3\},\{1\}\},\{\{1,2,3\}\}\big)\Big). \end{aligned}$$

Example 2 (Four Samples) When there are four samples, we have two f-sequences and eighteen c-sequences. We denote the f-sequences by $f^{\rangle}$ and $f^{\wedge}$. We can apply (8) to $\mathcal{C}_4$ and find that 12 c-sequences map to $f^{\rangle}$ and 6 map to $f^{\wedge}$. They are depicted in Fig. 2 as f-matrices.

In the Hasse diagram of $\mathbb{F}_n$ (see Fig. 3), the states $f_1, \ldots, f_n$ 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 transition probabilities of $\{F^{\uparrow}(k)\}_{k\in[n]_-}$ and $\{F^{\downarrow}(k)\}_{k\in[n]_+}$. A similar diagram for $n = 7$ appears in context of a breadth-first counting algorithm that sets the stage for an asymptotic enumerative study of the size of $\mathcal{F}_n$ (Erdös and Guy 1975, Fig. 1).

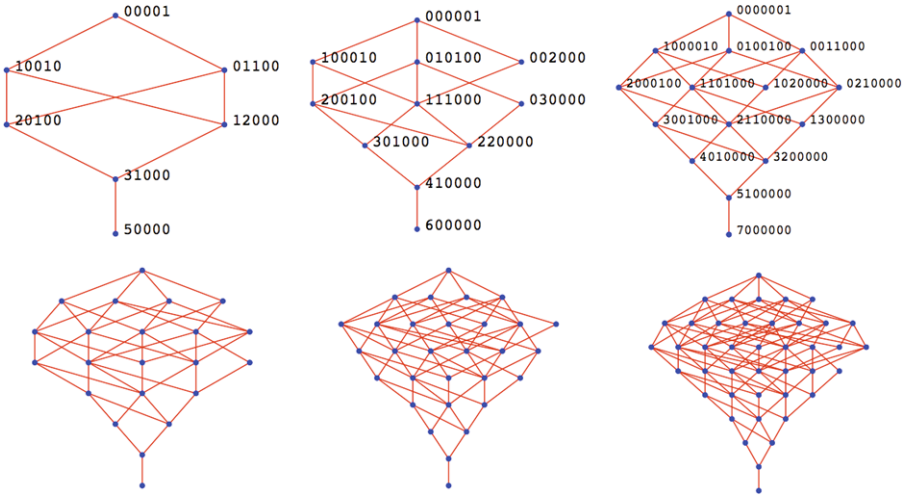

Fig. 3 Hasse diagrams of the state transition diagrams of the backward and forward Markov chains, $\{F^{\uparrow}(k)\}_{k\in[n]_-}$ and $\{F^{\downarrow}(k)\}_{k\in[n]_+}$, respectively, on $\mathbb{F}_n$ for $n = 5, 6, 7$ on *top row* with labeled states and $n = 8, 9, 10$ in *bottom row*

2.4 Exponentially Growing Population

So 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:

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

where $N(0)$ is the current population size. Let $A_{k:n} := \sum_{j=k}^{n} A_j$ denote the partial sum. 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 (Tavaré 1984):

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

2.5 Mutation Models

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. Figure 5 shows a coalescent tree for a sample of four individuals. In neutral models considered here under parameter $\phi = (\phi_1, \phi_2) \in \boldsymbol{\Phi}$, mutations are independently super-imposed upon the coalescent trees at each site according to a model of mutation for a specific biological marker with two or more states. The basic idea involves mutating the sampled or given state at an ancestral node to a possibly different state at the descendent node with a probability that depends on the mutation model and the lineage length between the two nodes. The two basic types of mutation models in population genetics are briefly summarized below.

2.5.1 *Infinitely-Many-Sites Models*

Under the infinitely-many-sites (IMS) model (Watterson 1975), independent mutations are super-imposed on the coalescent tree ${}^{c}t$ at each site according to a homogeneous Poisson process at rate $\phi_1 l_\bullet$, where $\phi_1 := 4N_e\mu$, $l_\bullet$ is the total size of the tree, N_e is the effective population size, 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.

2.5.2 *Finitely-Many-States Models*

There are several finitely-many-states models. A continuous-time Markov chain over finitely many states is used to model mutation from one state to another at each site. For example, over the nucleotide state space, a simple symmetric model (Jukes and Cantor 1969) allows transitions between any two distinct states at rate $\mu/3$. Mutations are modeled independently across sites over a given coalescent tree ${}^{c}t$ whose lineage lengths are in units of $4N_e$.

3 *n*-Coalescent Experiments

We give the statistical formalities needed to graphically frame our n-coalescent statistical experiments. Recall that a statistical experiment $(\mathcal{X}_n^m, \sigma(\mathcal{X}_n^m), \mathcal{P}_{\boldsymbol{\Phi}})$ is the ordered triple consisting of the sample space $\mathcal{X}_n^m$, a sigma-algebra over the sample space $\sigma(\mathcal{X}_n^m)$ and an identifiable $\boldsymbol{\Phi}$-indexed family of probability measures $\mathcal{P}_{\boldsymbol{\Phi}}$, i.e., $\boldsymbol{\Phi} \ni \phi \mapsto P_\phi \in \mathcal{P}_{\boldsymbol{\Phi}}$, over the sample space, such that, $P_\phi := P(x|\phi) \in \mathcal{P}_{\boldsymbol{\Phi}}$ for each $\phi \in \boldsymbol{\Phi}$. Our samples spaces $\mathcal{V}_n^m$ and $\mathcal{X}_n^m$ are finite and, therefore, P_ϕ's are dominated by the counting measure. Our continuous parameter space in this study is two-dimensional, i.e., $\boldsymbol{\Phi} := (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) \subset \mathbb{R}_+^2$. The first parameter ϕ_1 is the per-locus mutation rate scaled by the effective population size and is often denoted by θ in population genetics literature. The second parameter ϕ_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. We are interested in approximately sufficient statistics (Cam 1964) for the purpose of computational efficiency. Recall that a statistic $T_{\alpha,\beta}(z_\alpha) = z_\beta : \mathcal{Z}_\alpha \to \mathcal{Z}_\beta$ is sufficient for the experiment $\mathcal{X}_\alpha = (\mathcal{Z}_\alpha, \sigma(\mathcal{Z}_\alpha), \mathcal{P}_{\boldsymbol{\Phi}})$, provided:

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

for any $\phi \in \boldsymbol{\Phi}$. 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 $\phi \in \boldsymbol{\Phi}$, we get Bayes sufficiency in the Kolmogorov sense (Kolmogorov 1942), in terms of the following posterior identity:

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

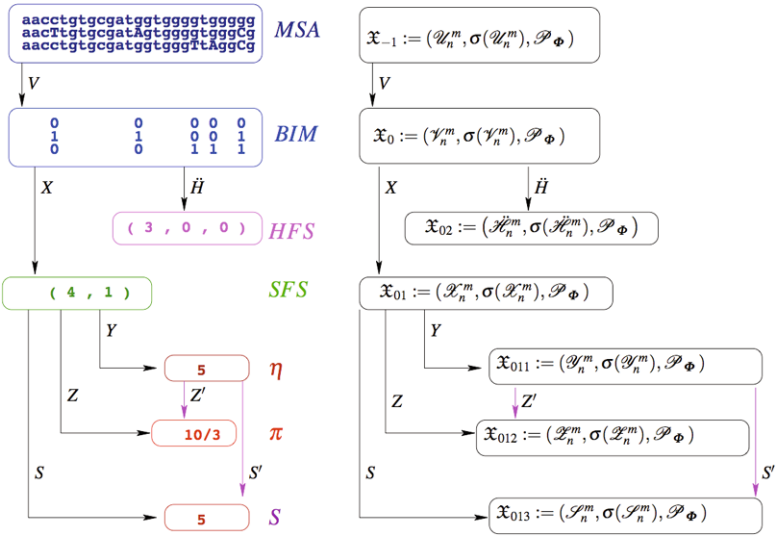

Fig. 4 An n-coalescent experiments graph. An observed multiple sequence alignment of the mother experiment and its offspring are shown on the *left*. The corresponding formalities are shown on the *right*

The fundamental experiment of this study is $\mathfrak{X}_{01} := (\mathcal{X}_n^m, \sigma(\mathcal{X}_n^m), \mathcal{P}_{\boldsymbol{\Phi}})$ at the resolution of SFS. We also pursue $\mathfrak{X}_0 := (\mathcal{V}_n^m, \sigma(\mathcal{V}_n^m), \mathcal{P}_{\boldsymbol{\Phi}})$ using existing methods for comparison. The other experiment nodes in the experiments graph of Fig. POEG are included to decision-theoretically unify various classical population genetic experiments. They include $(\ddot{\mathcal{H}}_n^m, \sigma(\ddot{\mathcal{H}}_n^m), \mathcal{P}_{\boldsymbol{\Phi}})$ that is based on the haplotype frequency spectrum or HFS $\ddot{H}$ (Ewens 1972, 1974), and the three linear subexperiments of $\mathfrak{X}_{01}$, namely, $\mathfrak{X}_{011} := (\mathcal{Y}_n^m, \sigma(\mathcal{Y}_n^m), \mathcal{P}_{\boldsymbol{\Phi}})$ for the folded site frequency spectrum or FSFS Y, $\mathfrak{X}_{012} := (\mathcal{Z}_n^m, \sigma(\mathcal{Z}_n^m), \mathcal{P}_{\boldsymbol{\Phi}})$ for the heterozygosity Z and $\mathfrak{X}_{013} := (\mathcal{S}_n^m, \sigma(\mathcal{S}_n^m), \mathcal{P}_{\boldsymbol{\Phi}})$ for the number of segregating sites $S = \sum_{i=1}^{n-1} x_i$. Using Markov bases, we approach the Tajima's D product experiment of $\mathfrak{X}_{012}$ and $\mathfrak{X}_{013}$.

3.1 Multiple Sequence Alignment

The data u_o 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. The mutation model is typically a reversible Markov model on the nucleotide state space $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ under the assumption of independence across sites. The conditional probability $P(u_o|\phi)$ that is proportional to the likelihood of ϕ is computed by integrating over all ancestral nucleotide states using a product-sum algorithm (Felsenstein 1981) for each coalescent tree ${}^{c}t$ in the coalescent tree space ${}^{\mathcal{C}_n}\mathbb{T}_n$ that is distributed according to ϕ.

Exact maximum likelihood estimation (e.g., Yang 2000; Hosten et al. 2005; Casanellas et al. 2005) as well as exact posterior sampling (Sainudiin and York 2009) is only feasible for small sample sizes ($n \leq 4$). The standard approach is to rely on Monte Carlo Markov chain (MCMC) algorithms (Metropolis et al. 1953; Hastings 1970) to obtain dependent samples from the posterior under the assumption

that the algorithm has converged to the desired stationary distribution. Unfortunately, there are no proven bounds for the burn-in period and thin-out rate that are needed to obtain Monte Carlo standard errors (Jones and Hobert 2001) from the MCMC samples. Thus, there is no guarantee that an MCMC sampler is indeed close to the desired stationary distribution over ${}^{\mathcal{C}_n}\mathbb{T}_n$ (Mossel and Vigoda 2005, 2006). Moreover, polymorphic sites are typically biallelic in human population genomic data. Thus, one need not have a finite state Markov model of mutations to explain most of the observed data patterns and can thereby circumvent the computational demands on evaluating the likelihood at the finest resolution of the MSA.

3.2 Binary Incidence Matrix

We assume the ancestral nucleotides are known, and at most one derived nucleotide occurs at each site among the sampled sequences (such biallelic data is common and sites showing ancestral and derived characters are commonly referred to as *single nucleotide polymorphisms* or SNPs). Then from the aligned sequence data u, we obtain a BIM $v \in \mathcal{V}_n^m := \{0,1\}^{n\times m}$ by replacing all ancestral states with 0 and derived states with 1.

BIM data is modeled by superimposing Watterson's infinitely-many-sites (IMS) model of mutation (Watterson 1975) over an n-coalescent sample genealogy (Kingman 1982a, 1982b). We can conduct inference on the basis of the observed *binary incidence matrix* or *BIM* v using existing importance sampling methods (e.g., Griffiths and Tavare 1994, 1996; Bahlo and Griffiths 1996; Stephens and Donnelly 2000; Slatkin 2002; Iorio and Griffiths 2004; Birkner and Blath 2008). In this study, we are not interested in inference on the basis of the observed BIM at a single locus, but instead on its SFS, a further summary of BIM.

3.3 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 ith entry of x records at how many sites exactly i sequences in u show the derived state. We assume that no site displays only the derived state. Thus, x has only $n-1$ entries. Figure 5 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. Next, we describe the basic probability models required to compute the likelihood of SFS.

3.3.1 Inference under the Unlabeled n-Coalescent

For a given coalescent tree ${}^c t \in {}^{\mathcal{C}_n}\mathbb{T}_n$, let the map:

$$L\left({}^c t\right) = l := (l_1, l_2, \ldots, l_{n-1}) : {}^{\mathcal{C}_n}\mathbb{T}_n \to \mathcal{L}_n := \mathbb{R}_+^{n-1} \tag{16}$$

compress the tree ${}^c t$ into the $n-1$ lineage lengths that could lead to singleton, doubleton, ..., and "$(n-1)$-ton" observations of mutationally derived states, respectively,

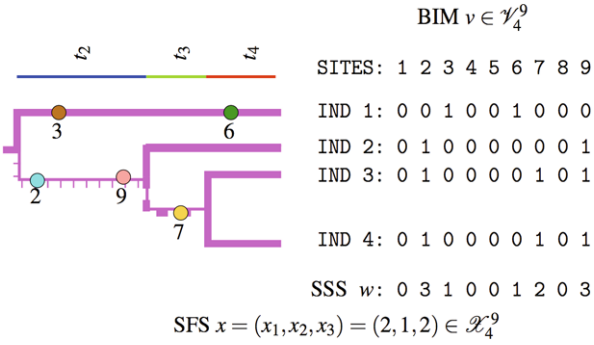

Fig. 5 At most, one mutation per site under the infinitely-many-sites model are superimposed as a homogeneous Poisson process upon the realization of identical coalescent trees at nine homologous SITES labeled $\{1, 2, \ldots, 9\}$ that constitute a non-recombining locus from four INDividuals labeled $\{1, 2, 3, 4\}$

i.e., l_i is the length of all the lineages in ${}^{c}t$ that subtend i samples or leaves. For example in Fig. 5, (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 at each one of the m sites at our locus. Now, let $\bar{l}_i := l_i / l_\bullet$ be the relative length of lineages that subtend i leaves at each site. Now, define $\bar{l} := (\bar{l}_1, \bar{l}_2, \ldots, \bar{l}_{n-1}) \in \Delta_{n-2}$, 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:

$$P\left(x|\phi, {}^{c}t\right) = P(x|\phi, l) = e^{-\phi_1 m l_\bullet} (\phi_1 m l_\bullet)^s \prod_{i=1}^{n-1} \bar{l}_i^{x_i} \Bigg/ \prod_{i=1}^{n-1} x_i!, \tag{17}$$

where $s = \sum_{i=1}^{n-1} x_i$ is the number of segregating sites. The distribution on ${}^{\mathcal{C}_n}\mathbb{T}_n$ is given by the ϕ_2-indexed n-coalescent approximation of the sample genealogy in an exponentially growing Wright–Fisher model. This distribution on ${}^{\mathcal{C}_n}\mathbb{T}_n$ in turn determines the distribution of the random vector L on $\mathcal{L}_n$. We employ the appropriate lumped Markov process to efficiently obtain $P(\phi|x)$ as per Remark 2.

Remark 2 Kemeney 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.*"

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 nth Bell number for a range of sample sizes is tabulated below (see Table 1).

Table 1 Cardinalities of the state spaces

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

Using the unlabeled n-coalescent, we can directly prescribe the ϕ-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 ϕ and an f-sequence $f \in \mathcal{F}_n$. Recall $P(x|\phi, {}^c t) = P(x|\phi, l)$, where $l = L({}^c t)$, as in (17). We show that l is determined by the f-matrix $\mathbf{f} = \mathbf{F}(f)$ of the f-sequence $f = \underline{\mathscr{F}}(c)$ of the c-sequence c and the epoch-times vector t of the coalescent tree ${}^c t$.

Proposition 5 (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, $f = \underline{\mathscr{F}}(c)$ be its f-sequence, $\mathbf{f} = \mathbf{F}(f)$ be its f-matrix and $t = (t_2, t_3, \ldots, t_n) \in (0, \infty)^{n-1}$ be its epoch times as a column vector and its transpose t^{T} be the corresponding row vector. Then $L({}^c t) = l$ of (16) is given by the following matrix multiplications*:

$$l = t^{\mathrm{T}}\mathbf{f} = \left(\sum_{i=2}^{n} t_i f_{i,1}, \sum_{i=2}^{n-1} t_i f_{i,2}, \ldots, \sum_{i=2}^{2} t_i f_{i,n-1} \right). \tag{18}$$

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

$$\begin{aligned} P\left(x|\phi, {}^c t\right) &= P(x|\phi, l) = P\left(x|\phi, t^{\mathrm{T}}\mathbf{f}\right) \\ &= \frac{1}{\prod_{i=1}^{n-1} x_i!} \exp\left(-\phi_1 m \sum_{j=1}^{n-1} \sum_{i=2}^{n+1-j} t_i f_{i,j} \right) \left(\phi_1 m \sum_{j=1}^{n-1} \sum_{i=2}^{n+1-j} t_i f_{i,j} \right)^{\sum_{i=1}^{n-1} x_i} \\ &\quad \times \prod_{i=1}^{n-1} \left(\sum_{i=2}^{n+1-j} t_i f_{i,j} \left(\sum_{j=1}^{n-1} \sum_{i=2}^{n+1-j} t_i f_{i,j} \right)^{-1} \right)^{x_i}. \end{aligned} \tag{19}$$

Proof The proof of (18) is merely a consequence of the encoding of f as the matrix $\mathbf{f}$ and (19) follows from (18) and (17). □

The computation of l from t and $\mathbf{f}$ requires at most $n^2 - 2n + 1$ multiplications and additions 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 ϕ_1 per site
2. sample size n
3. number of sites m at the locus

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 $t \sim T = (T_2, T_3, \ldots, T_n) \sim \bigotimes_{i=2}^{n} \binom{i}{2} e^{-\binom{i}{2} t_i}$, or as desired from $\mathbb{R}_+^{n-1}$
5: draw x from the (f,t)-dependent Poisson-Multinomial distribution of (19)
6: **return**: x

Note that Algorithm 1 is quite general since the only restriction on t in step 4 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 provided the c-sequence c and its f-sequence $f = \underline{\mathscr{F}}(c)$ 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 vector t.

Next we study one f-sequence in detail as it is an interesting extreme case that will resurface in the sequel.

Example 3 (Completely unbalanced tree) Let the f-sequence $f^{\curlywedge} \in \mathcal{F}_n$ denote that of the completely unbalanced tree. Its probability based on (9) and (10) are:

$$f^{\curlywedge} := \left(f_1^{\curlywedge}, f_2^{\curlywedge}, \ldots, f_n^{\curlywedge}\right), \quad \text{where } f_i^{\curlywedge} = (i-1)\, e_1 + e_{(n-i+1)}, \tag{20}$$

$$P\left(f^{\curlywedge}\right) = \prod_{i=n}^{2} P\left(f_{i-1}^{\curlywedge} | f_i^{\curlywedge}\right) = \prod_{i=n-1}^{2} \frac{(i-1)1}{i(i-1)/2} = \frac{2^{n-2}}{(n-1)!}. \tag{21}$$

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

The posterior distribution $P(\phi|x) \propto P(x|\phi)P(\phi)$ over $\boldsymbol{\Phi}$ 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 $\underline{\mathscr{F}}^{-1}(f)$, where f is the f-sequence corresponding to the c-sequence of ${}^{c}t$. For samples of size $2 \le n \le 3$, there is only one f-sequence in $\mathcal{F}_n$. For samples with $n \ge 4$, 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 $\mathbf{2}^{\{1,2,\ldots,n-1\}}$, the power set of $\{1,2,\ldots,n-1\}$:

$$X^{\circledast}(x) = x^{\circledast} := \left(x_1^{\circledast}, \ldots, x_{n-1}^{\circledast}\right) := \left(\mathbb{1}_{\mathbb{N}}(x_1), \ldots, \mathbb{1}_{\mathbb{N}}(x_{n-1})\right) : \mathcal{X}_n^m \to \{0,1\}^{n-1}.$$

If $x_h^{\circledast} = 1$ then the hth entry of the SFS x is at least one, i.e., $x_h > 0$. Thus, $X^{\circledast}(x) = x^{\circledast}$ encodes the presence or absence of at least one site's ancestral lineage that has

been hit by a mutation while subtending h samples, where $h \in \{1, 2, \ldots, n-1\}$. Next, consider the following two sets of f-sequences:

$$F_n\big(x^{\circledast}\big) := \bigcup_{\{h:\, x_h^{\circledast}=1\}} \left\{ f \in \mathcal{F}_n : \sum_{i=1}^{n} f_{i,h} = 0 \right\}, \qquad \complement F_n\big(x^{\circledast}\big) := \mathcal{F}_n \setminus F_n\big(x^{\circledast}\big). \tag{22}$$

The set of f-sequences $F_n(x^{\circledast})$ and its complement $\complement F_n(x^{\circledast})$ play a fundamental role in inference from an SFS x and its $X^{\circledast} = x^{\circledast}$. Note that when an SFS x has none of the x_i's equaling 0, then its $x^{\circledast} = (1, 1, \ldots, 1)$ and $\complement F_n(x^{\circledast})$ only contains the f-sequence corresponding to the completely unbalanced tree $f^{\rangle}$ given by (20). At the other extreme, when an SFS x has all its x_i's equaling 0 with $x^{\circledast} = (0, 0, \ldots, 0)$, we are unable to discriminate among f-sequences since $\complement F_n(x^{\circledast}) = \mathcal{F}_n$. Thus,

$$\complement F_n(0, 0, \ldots, 0) = \mathcal{F}_n \quad \text{and} \quad \complement F_n(1, 1, \ldots, 1) = \big\{ f^{\rangle} \big\}. \tag{23}$$

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

Proposition 6 (Likelihood of SFS) *For any $t \in (0, \infty)^{n-1}$ and any $x \in \mathcal{X}_n^m$ with $x^{\circledast} = X^{\circledast}(x)$,*

$$\text{If } f \in F_n\big(x^{\circledast}\big) \text{ and } l = t^{\mathrm{T}} \cdot \mathbf{F}(f) \text{ then } \prod_{i=1}^{n-1} \bar{l}_i^{x_i} = 0. \tag{24}$$

Therefore, the likelihood of SFS x is proportional to

$$\begin{aligned} P(x|\phi) = \frac{1}{\prod_{i=1}^{n-1} x_i!} \sum_{f \in \complement F_n(x^{\circledast})} P(f) \Bigg(\int_{t \in (0,\infty)^{n-1}} \Bigg(\exp\Bigg(-\phi_1 m \sum_{j=1}^{n-1} \sum_{i=2}^{n+1-j} t_i f_{i,j} \Bigg) \\ \times \Bigg(\phi_1 m \sum_{j=1}^{n-1} \sum_{i=2}^{n+1-j} t_i f_{i,j} \Bigg)^{\sum_{i=1}^{n-1} x_i} \\ \times \prod_{i=1}^{n-1} \Bigg(\sum_{i=2}^{n+1-j} t_i f_{i,j} \Bigg(\sum_{j=1}^{n-1} \sum_{i=2}^{n+1-j} t_i f_{i,j} \Bigg)^{-1} \Bigg)^{x_i} \Bigg) \, dP(t|\phi) \Bigg). \end{aligned} \tag{25}$$

Proof We first prove the implication in (24). Given any $t \in (0, \infty)^{n-1}$ and any $x \in \mathcal{X}_n^m$ with $x^{\circledast} = X^{\circledast}(x)$, let $f \in F_n(x^{\circledast})$. First, suppose $x_h^{\circledast} = 0$ for every $h \in \{1, 2, \ldots, n-1\}$, then $F_n(x^{\circledast}) = \emptyset$ and we have nothing to prove. Now, suppose there exists some h such that $x_h^{\circledast} = 1$, or equivalently $x_h > 0$, then by the constructive definition of $F_n(x^{\circledast})$, we have that for any $f \in F_n(x^{\circledast})$ $\sum_{i=1}^{n} f_{i,h} = 0$, which implies that $f_{i,h} = 0$ for every $i \in \{1, 2, \ldots, n\}$ since $f_{i,j} \geq 0$. Therefore, by applying this implication to the expression for l_h in Proposition 5, 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 (24) is a consequence of $\bar{l}_h^{x_h} = (l_h / l_\bullet)^{x_h} = 0^{x_h} = 0$.

Next, we prove (25). For simplicity, we abuse notation and write $P(\cdot)$ to denote the probability as well as the probability density under the appropriate dominating measure. 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 $P(x|\phi) = P(x, \phi)/P(\phi)$:

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

since by independence of f and (ϕ, 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

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

Substituting for $P(x|\phi, l = t^{\mathrm{T}} \cdot \mathbf{f})$ from Proposition 5 and only summing over $f \in \mathcal{CF}_n(x^{\circledast})$ with nonzero probability $P(x|\phi, l = t^{\mathrm{T}} \cdot \mathbf{f})$, we get the discrete sum weighted by integrals on $\mathbb{T}_n := (0, \infty)^{n-1}$, the required equality in (25). □

Next, we devise an algorithm to estimate $P(x|\phi)$, the probability of an observed SFS x given a parameter ϕ. This is accomplished by constructing a Markov chain $\{F^{\lfloor x^{\circledast}}(k)\}_{k \in [n]_+}$ on the state space $\mathbb{F}_n^{x^{\circledast}} \subset \mathbb{F}_n \times \{0, 1\}^{n-1}$ such that every sequence of states visited by this chain yields a probable f-sequences f for the observed SFS x, i.e., $f \in \mathcal{CF}_n(x^{\circledast})$. In this paper, we focus on small $n \in \{4, 5, \ldots, 10\}$ and exhaustively sum over all $f \in \mathcal{CF}_n(x^{\circledast})$ that are unique sequential realizations of $\{F^{\lfloor x^{\circledast}}(k)\}_{k \in [n]_+}$. The maximal number of such f-sequences is

$$\max_{x^{\circledast} \in \{0,1\}^{n-1}} \left|\mathcal{CF}_n\big(x^{\circledast}\big)\right| = \left|\mathcal{CF}_n\big((0, 0, \ldots, 0)\big)\right| = |\mathcal{F}_n|\,.$$

A breadth-first search on the transition graph of $\{F^{\lfloor x^{\circledast}}(k)\}_{k \in [n]_+}$ revealed that

$$\begin{aligned} |\mathcal{F}_n| &= 2, 4, 11, 33, 116, 435, 1832, \ldots, 6237505, \\ &\text{as } \; n = 4, 5, 6, 7, 8, 9, 10, \ldots, 15, \end{aligned}$$

respectively. Our computations are in agreement with similar numerical calculations of $|\mathcal{F}_n|$ in Erdös and Guy (1975, Sect. 2). This $x^{\circledast}$-indexed family of 2^{n-1} Markov

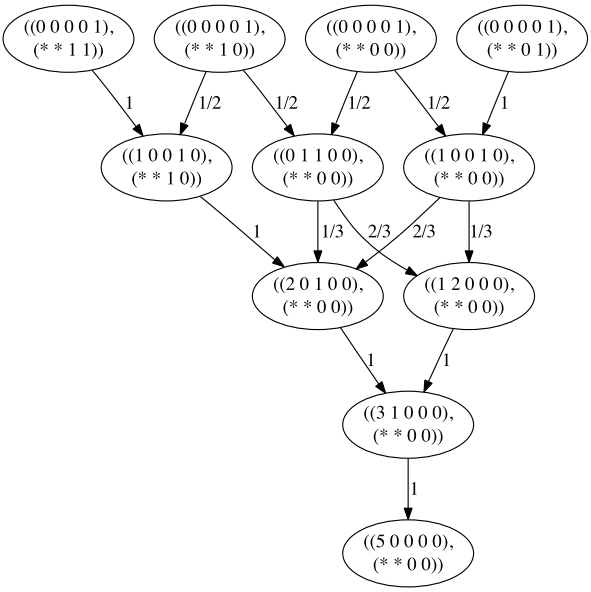

Fig. 6 Transition diagram of $\{F^{\lfloor x^{\circledast}}(k)\}_{k\in[5]_+}$ over states in $\mathbb{F}_n^{x^{\circledast}}$. The simplified diagram replaces the states that do not affect the transitions, namely, $x_1^{\circledast}$ and $x_2^{\circledast}$, with $* \in \{0, 1\}$

chains $\{F^{\lfloor x^{\circledast}}(k)\}_{k\in[n]_+}$ 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., Duflo 1997, Sect. 7.3) over the state space $\mathbb{F}_n$ with control space $\{0,1\}^{n-1}$ that can produce the desired f-sequences in $\complement_{F_n}(x^{\circledast})$.

Optimal importance sampling by using the sequential realizations of $\{F^{\lfloor x^{\circledast}}(k)\}_{k\in[n]_+}$ and its continuous time variant as a proposal distribution in order to get the Monte Carlo estimate of $P(x|\phi)$ for larger n 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 ϕ-indexed measures on $\mathcal{C}_n\mathbb{T}_n$.

Proposition 7 (A Proposal over $\complement_{F_n}(x^{\circledast})$) *For a given SFS $x \in \mathcal{X}_n^m$ and $X^{\circledast}(x) = x^{\circledast} \in \{0,1\}^{n-1}$, consider the discrete time Markov chain $\{F^{\lfloor x^{\circledast}}(k)\}_{k\in[n]_+}$ over the state space of ordered pairs $(f_{i'}, z_{i'}) \in \mathbb{F}_n^{x^{\circledast}} \subset \mathbb{F}_n \times \{0,1\}^{n-1}$, with the initial state given by $(f_1, x^{\circledast}) = ((0,0,\ldots,1), x^{\circledast})$, the transition probabilities obtained by a controlled reweighing of the transition probabilities of $\{F^{\downarrow}(k)\}_{k\in[n]_+}$ over $\mathbb{F}_n$ as follows*:

$$P\big((f_{i'}, z_{i'})|(f_i, z_i)\big) = \begin{cases} P(f_{i'}|f_i)/\Sigma(f_i, z_i) & :if \quad (f_i, z_i) \prec_{f,z} (f_{i'}, z_{i'}), \\ 0 & :otherwise, \end{cases} \tag{26}$$

where

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

$$\Xi(f_i, z_i) \;:=\; \big\{(j,k) : f_{i,j+k} > 0,\ 1 \le j \le \widehat{j} \le k \le j+k-1\big\},$$

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

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

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

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

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

Then $\mathcal{F}_n^{x^{\circledast}} = \complement_{F_n}(x^{\circledast})$.

Proof We will prove that $\mathcal{F}_n^{x^{\circledast}} = \complement_{F_n}(x^{\circledast})$ for three cases after noting that the orthonormal 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 $x^{\circledast} = (0,0,\ldots,0)$. Since $\complement_{F_n}(x^{\circledast}) = \mathcal{F}_n$ by (23), we need to show that $\mathcal{F}_n^{x^{\circledast}} = \mathcal{F}_n$. Initially, at time step 1,

$$F^{\downarrow x^{\circledast}}(1) = (f_1, z_1) = \left(f_1, x^{\circledast}\right) = \left((0,0,\ldots,0,1),(0,0,\ldots,0)\right).$$

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

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

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

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 $x^{\circledast} = (0,0,\ldots,0)$ our fully uncontrolled Markov chain $\{F^{\downarrow x^{\circledast}}(k)\}_{k\in[n]_+}$ visits states in $\mathbb{F}_n$ in a manner identical to the Markov chain $\{F^{\downarrow}(k)\}_{k\in[n]_+}$ over $\mathbb{F}_n$. Therefore, $\mathcal{F}_n^{x^{\circledast}} = \mathcal{F}_n = \complement_{F_n}(x^{\circledast})$ when $x^{\circledast} = (0,0,\ldots,0)$.

Case 2: Suppose $x^{\circledast} = (1,1,\ldots,1)$. Since $\complement_{F_n}(x^{\circledast}) = \{f^{\searrow}\}$ by (23), we need to show that $\mathcal{F}_n^{x^{\circledast}} = \{f^{\searrow}\}$. Initially, at time step 1,

$$F^{\downarrow x^{\circledast}}(1) = (f_1, z_1) = \left(f_1, x^{\circledast}\right) = \left((0,0,\ldots,0,1),(1,1,\ldots,1,1)\right)$$

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

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

Thus, the only state that is immediately preceded by (f_1, z_1) is our next state $(f_2, z_2) = (f_1 - e_n + e_1 + e_{n-1}, z_1 - \mathbb{1}_{\{1\}}(z_{1,1})\, e_1 - \mathbb{1}_{\{1\}}(z_{i,n-1})\, e_{n-1})$ with probability 1 due to the equality of the numerator and denominator in (26):

$$(f_1, z_1) \prec_{f,z} (f_2, z_2) = \big((1, 0, \ldots, 1, 0), (0, 1, \ldots, 1, 0)\big) = F^{\lfloor x^{\circledast}}(2).$$

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

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

By (20), $f_{i+1} = e_{n-i} + i e_1 = f_{i+1}^{\rangle}$ and we get the desired $f^{\rangle} = (f_n^{\rangle}, f_{n-1}^{\rangle}, \ldots, f_1^{\rangle})$ in the forward direction as the only realization over $\mathbb{F}_n$ of our fully controlled Markov chain $\{F^{\lfloor x^{\circledast}}(k)\}_{k \in [n]_+}$. Therefore, $\mathcal{F}_n^{x^{\circledast}} = \{f^{\rangle}\} = \mathsf{C}_{\mathcal{F}_n}(x^{\circledast})$ when $x^{\circledast} = (1, 1, \ldots, 1)$.

Case 3: Now, suppose $x^{\circledast} \in \{0,1\}^{n-1} \setminus \{(0, 0, \ldots, 0), (1, 1, \ldots, 1)\}$. First, we will show that $f \in \mathcal{F}_n^{x^{\circledast}}$ implies that $f \in \mathsf{C}_{\mathcal{F}_n}(x^{\circledast})$ or equivalently that $f \notin \mathcal{F}_n(x^{\circledast})$. We will prove by contradiction. Assume $f \in \mathcal{F}_n^{x^{\circledast}}$. Suppose that $f \in \mathcal{F}_n(x^{\circledast})$. Then by (22), there exists an h with $x_h^{\circledast} = 1$ such that $\sum_{i=1}^{n} f_{i,h} = 0$. Since $\sum_{i=1}^{n} f_{i,1} > 0$ and $\sum_{i=1}^{n} f_{i,2} > 0$ for every $f \in \mathcal{F}_n$, with $n > 2$, $h \in \{3, 4, \ldots, n-1\}$. 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 $f = (f_1, f_2, \ldots, f_n)$ over $\mathbb{F}_n$ by $\{F^{\lfloor x^{\circledast}}(k)\}_{[n]_+}$. This contradicts our assumption that $f \in \mathcal{F}_n^{x^{\circledast}}$ 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 $\widehat{j}$. So, our supposition that $f \in \mathcal{F}_n(x^{\circledast})$ is false. Therefore, if $f \in \mathcal{F}_n^{x^{\circledast}}$ then $f \in \mathsf{C}_{\mathcal{F}_n}(x^{\circledast})$. Next, we will show $f \in \mathsf{C}_{\mathcal{F}_n}(x^{\circledast})$ implies that $f \in \mathcal{F}_n^{x^{\circledast}}$. Assume that $f \in \mathsf{C}_{\mathcal{F}_n}(x^{\circledast})$, then $\sum_{i=1}^{n} f_{i,h} > 0$ for every $h \in \{h : x_h^{\circledast} = 1\}$ by (22). This means that for each h with $x_h^{\circledast} = 1$ 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 $\max\{\ell : z_{i,\ell} = 1\} = h$, $h \in \{h : x_h^{\circledast} = 1\}$, in the definition of $\widehat{j}$, this f can be sequentially realized over $\mathbb{F}_n$ by $\{F^{\lfloor x^{\circledast}}(k)\}_{[n]_+}$. Therefore, if $f \in \mathsf{C}_{\mathcal{F}_n}(x^{\circledast})$ then $f \in \mathcal{F}_n^{x^{\circledast}}$. □

Thus, given ϕ_1 and an $x^{\circledast}$, we can efficiently propose SFS samples from $\mathcal{X}_n^m$, such that the underlying f-sequence f belongs to $\mathsf{C}_{\mathcal{F}_n}(x^{\circledast})$, using Algorithm 2. Note, however, that a further straightforward importance sampling step using (26) and (12) is needed to obtain SFS samples that are distributed over $\mathcal{X}_n^m$ according to the unlabeled n-coalescent over $\mathsf{C}_{\mathcal{F}_n}(x^{\circledast})$.

3.4 Linear Experiments of the Site Frequency Spectrum

We describe a method to obtain the conditional probability $P(r|\phi, {}^{c}t)$, where $r = \mathbf{R}x$ is a set of classical population genetic statistics that are linear combinations of the site

Algorithm 2 SFS Proposal under an $x^{\circledast}$-controlled unlabeled n-coalescent

1: **input**:

1. scaled mutation rate ϕ_1 per site
2. number of sites m at the locus
3. observed $x^{\circledast}$ (note that sample size $n = |x^{\circledast}| + 1$)

2: **output**: an SFS sample x such that the underlying f-sequence $f \in \mathsf{C}_{\mathsf{F}_n}(x^{\circledast})$
3: generate an f-sequence f under $\{F^{|x^{\circledast}|}(k)\}_{k\in[n]_+}$
4: draw $t \sim T = (T_2, T_3, \ldots, T_n) \sim \bigotimes_{i=2}^{n} \binom{i}{2} e^{-\binom{i}{2}t_i}$, or as desired from $\mathbb{R}_+^{n-1}$
5: draw x from the (f,t)-dependent Poisson-multinomial distribution of (19)
6: **return**: x

frequency spectrum x, ϕ is the vector of parameters in the population genetic model and ${}^{c}t$ is the underlying coalescent tree upon which mutations are superimposed to obtain the data. The conditional probability is obtained by an appropriate integration over

$$\mathbf{R}^{-1}(r) := \left\{x : x \in \mathbb{Z}_+^{n-1}, \mathbf{R}x = r\right\}.$$

$\mathbf{R}^{-1}(r)$ is called a *fiber*.

We want to compute $P(r|\phi)$, since the posterior distribution of interest is $P(\phi|r) \propto P(r|\phi)P(\phi)$. Furthermore, we assume a uniform prior over a biologically sensible grid of ϕ values and evaluate $P(r|\phi)$ over each ϕ in our grid. More precisely, we have

$$P\left(r|\phi, {}^{c}t\right) = P\left(r|\phi, l = L\left({}^{c}t\right)\right) = \sum_{x\in\mathbf{R}^{-1}(r)} P(x|\phi, l), \tag{27}$$

$$P(r|\phi) = \int_{l\in\mathcal{L}_n} P(r|\phi, l)P(l|\phi) = \int_{l\in\mathcal{L}_n} \sum_{x\in\mathbf{R}^{-1}(r)} P(x|\phi, l)P(l|\phi). \tag{28}$$

We can approximate the two integrals in (28) by the finite Monte Carlo sums,

$$P(r|\phi) \approx \frac{1}{N}\sum_{j=1}^{N}\frac{1}{M}\sum_{\substack{h=1,\\ x^{(h)}\in\mathbf{R}^{-1}(r)}}^{M} P\left(x^{(h)}\middle|\phi, l^{(j)}\right), \quad l^{(j)} \sim P(l|\phi). \tag{29}$$

The inner Monte Carlo sum approximates $\sum_x P(x|\phi, l)$ over M $x^{(h)}$'s in $\mathbf{R}^{-1}(r)$ and the outer Monte Carlo sum over N different $l^{(j)}$'s can be obtained from simulation under ϕ. Therefore, $P(\phi|r) \propto P(r|\phi)P(\phi)$

$$\approx \frac{1}{N}\sum_{j=1}^{N}\frac{1}{M}\sum_{\substack{h=1,\\ x^{(h)}\in\mathbf{R}^{-1}(r)}}^{M} P\left(x^{(h)}\middle|\phi, l^{(j)}\right), \quad l^{(j)} \sim P(l|\phi)P(\phi).$$

If $|\mathbf{R}^{-1}|$ is not too large, say less than a million, then we can do the inner summation exactly by a breadth-first traversal of an implicit graph representation of $\mathbf{R}^{-1}(r)$.

In general, the sum over $\mathbf{R}^{-1}(r)$ is accomplished by a Monte Carlo Markov chain on a graph representation of the state space $\mathbf{R}^{-1}(r)$ that guarantees irreducibility. This article is mainly concerned with the application of Markov bases to facilitate these integrations over $\mathbf{R}^{-1}(r)$. Although Markov bases were first introduced in the context of exact tests for contingency tables (Diaconis and Sturmfels 1998), we show in this article that they can also be used to obtain the posterior distribution $P(\phi|r_o)$ of various observed population genetic statistics r_o.

Definition 1 (Markov basis) Let $\mathbf{R}$ be a $q \times (n-1)$ integral matrix. Let $\mathcal{M}_{\mathbf{R}}$ be a finite subset of the intersection of the kernel of $\mathbf{R}$ and $\mathbb{Z}^{n-1}$. Consider the undirected graph $\mathcal{G}^r_{\mathbf{R}}$, such that (1) the nodes are all lattice points in $\mathbf{R}^{-1}(r)$ and (2) edges between a node x and a node y are present if and only if $x - y \in \mathcal{M}_{\mathbf{R}}$. If $\mathcal{G}^r_{\mathbf{R}}$ is connected for all r with $\mathcal{G}^r_{\mathbf{R}} \neq \emptyset$, then $\mathcal{M}_{\mathbf{R}}$ is called a Markov basis associated with the matrix $\mathbf{R}$. We refer to an $m := (m_1, \ldots, m_{(n-1)}) \in \mathcal{M}_{\mathbf{R}}$ as a move.

A Markov basis can be computed with computational commutative algebraic algorithms (Diaconis and Sturmfels 1998) implemented in algebraic software packages such as `Macaulay 2` (Grayson and Stillman 2004) and `4ti2` (Hemmecke et al. 2005). Monte Carlo Markov chains constructed with moves from $\mathcal{M}_{\mathbf{R}}$ are irreducible and can be made aperiodic, and are therefore ergodic on the finite state space $\mathbf{R}^{-1}(r)$. An ergodic Markov chain is essential to sample from some target distribution on $\mathbf{R}^{-1}(r)$ using Monte Carlo Markov chain (MCMC) methods.

3.4.1 Number of Segregating Sites

A classical statistic in population genetics is S, the *number of segregating sites* (Watterson 1975). It can be expressed as the sum of the components of the SFS x:

$$S(x) := \sum_{i=1}^{n-1} x_i = s : \mathcal{X}_n^m \to \mathcal{S}_n^m. \tag{30}$$

S is the statistic of the n-coalescent experiment $\mathfrak{X}_{013} := (\mathcal{S}_n^m, \sigma(\mathcal{S}_n^m), \mathcal{P}_{\boldsymbol{\Phi}})$. For some fixed sample size n at m homologous and at most bi-allelic sites, let the s-simplex $S^{-1}(s) = \{x \in \mathcal{X}_n^m : S(x) = s\}$ denote the set of SFS that have the same number of segregating sites s. The size of $S^{-1}(s)$ is given by the number of compositions of s by $n-1$ parts, i.e., $|S^{-1}(s)| = \binom{s+n-2}{s}$. The conditional probability of S is Poisson distributed with rate parameter given by the product of the total tree size $l_\bullet := \sum_{i=1}^{n-1} l_i$, number of sites m and the per-site scaled mutation rate parameter ϕ_1 in ϕ

$$\begin{aligned} P\left(S = s \middle| \phi, {}^c t\right) = P(S = s|\phi, l) &= \sum_{x \in S^{-1}(s)} P(x|\phi, l) \\ &= \sum_{x \in S^{-1}(s)} e^{-\phi_1 m l_\bullet} (\phi_1 m l_\bullet)^s \prod_{i=1}^{n-1} \bar{l}_i^{x_i} \left(\prod_{i=1}^{n-1} x_i! \right)^{-1} \\ &= e^{-\phi_1 m l_\bullet} (\phi_1 m l_\bullet)^s / s! \end{aligned}$$

3.4.2 Heterozygosity

Another classical summary statistic called *average heterozygosity* is also a symmetric linear combination of SFS x (Tajima 1989). We define *heterozygosity* $Z(x) = z$ and average pair-wise heterozygosity $\Pi(x) = \pi$ for the entire locus as follows:

$$Z(x) := \sum_{i=1}^{n-1} i(n-i)x_i, \qquad \Pi(x) := \frac{1}{\binom{n}{2}} Z(x). \tag{31}$$

Z is the statistic of the n-coalescent experiment $\mathfrak{X}_{012} := (\mathcal{Z}_n^m, \sigma(\mathcal{Z}_n^m), \mathcal{P}_{\boldsymbol{\phi}})$. For some fixed sample size n at m homologous and at most biallelic sites, consider the set of SFS that have the same heterozygosity z denoted by $Z^{-1}(z) = \{x \in \mathcal{X}_n^m : Z(x) = z\}$. This set is the intersection of a hyper-plane with $\mathcal{X}_n^m$. The conditional probability $P(Z|\phi, {}^c t) = P(\Pi|\phi, {}^c t) = P(Z = z|\phi, l)$ is

$$P(Z = z|\phi, l) = \sum_{x \in Z^{-1}(z)} P(x|\phi, l) = e^{-\phi_1 m l_\bullet} \sum_{x \in Z^{-1}(z)} (\phi_1 m l_\bullet)^{\sum_{i=1}^{n-1} x_i} \frac{\prod_{i=1}^{n-1} \bar{l}_i^{x_i}}{\prod_{i=1}^{n-1} x_i!}.$$

3.4.3 Tajima's D

Tajima's D statistic (Tajima 1989) for a locus only depends on the number of segregating sites of (30), average pair-wise heterozygosity of (31) and the sample size n, as follows:

$$D(x) := \frac{\Pi(x) - S(x)/d_1}{\sqrt{d_3 S(x) + d_4 S(x)(S(x) - 1)}}, \tag{32}$$

where $d_1 := \sum_{i=1}^{n-1} i^{-1}$, $d_2 := \sum_{i=1}^{n-1} i^{-2}$,

$$d_3 := \frac{n+1}{3d_1(n-1)} - \frac{1}{d_1^2}, \qquad d_4 := \frac{1}{d_1^2 + d_2}\left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{nd_1} + \frac{d_2}{d_1^2}\right).$$

Thus, Tajima's D is a statistic of $\mathfrak{X}_{012} \times \mathfrak{X}_{013}$, a product n-coalescent experiment. Let $r = (s, z)'$ for a given sample size n. Observe that fixing n and r also fixes the average heterozygosity π and Tajima's d. Next, we will see that inference based on s, π and d for a fixed sample size n depends on the kernel or null space of the matrix $\mathbf{R}$ given by

$$\mathbf{R} := \begin{pmatrix} 1 & \dots & 1 & \dots & 1 \\ 1(n-1) & \dots & i(n-i) & \dots & (n-1)(n-(n-1)) \end{pmatrix}.$$

The space of all possible SFS x for a given sample size n is the nonnegative integer lattice $\mathbb{Z}_+^{n-1}$. Let the intersection of $\{x : \mathbf{R}x = r\}$ with $\mathbb{Z}_+^{n-1}$ be the set:

$$\mathbf{R}^{-1}(r) := \left\{x \in \mathbb{Z}_+^{n-1} : \mathbf{R}x = r\right\}.$$

Since n is fixed, every SFS x in $\mathbf{R}^{-1}(r)$ has the same s, z, π, and d.

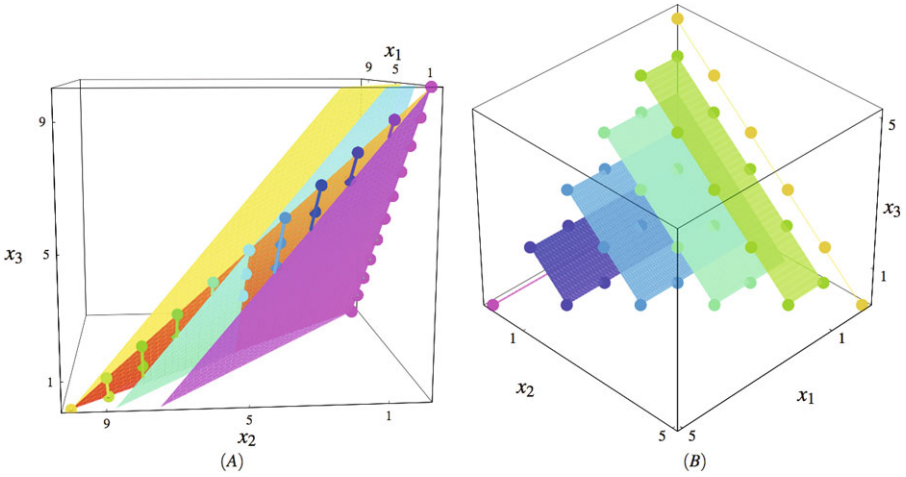

Fig. 7 **A** Polytopes containing $\mathbf{R}^{-1}((s, z)')$, where $z \in \{30, 31, \ldots, 40\}$, $n = 4$ and $s = 10$ are at the intersection of the s-simplex, $\mathbb{Z}_+^3$ and each of the z-simplexes. **B** Projected rectangular polytopes containing $\mathbf{R}^{-1}((s, z)')$, where $z \in \{20, 22, \ldots, 30\}$, $n = 5$ and $s = 5$ (see text)

When $n = 4$, we can visualize any SFS $x \in \mathbf{R}^{-1}(r)$ using Cartesian coordinates. Let $\mathbf{R}_1 = (1, 1, 1)$ and ${\mathbf{R}_1}^{-1}(s) := \{x \in \mathbb{Z}_+^3 : \sum_{i=1}^3 x_i = s\}$, the set of SFS with s segregating sites, be formed by the intersection of $\mathbb{Z}_+^3$ with the s-simplex given by $x_3 = s - x_1 - x_2$. Figure 7A shows ${\mathbf{R}_1}^{-1}(10)$, the set of 66 SFS with 10 segregating sites, as colored balls embedded in the orange s-simplex with $s = 10$. Similarly, with $\mathbf{R}_2 = (3, 4, 3)$, ${\mathbf{R}_2}^{-1}(z) := \{x \in \mathbb{Z}_+^3 : 3x_1 + 4x_2 + 3x_3 = z\}$ is the set of SFS at the intersection of $\mathbb{Z}_+^3$ with the z-simplex given by $x_3 = (z - 3x_1 - 4x_2)/3$. Figure 7A shows three z-simplexes for $z = 30, 35$ and 40, in hues of violet, turquoise, and yellow, respectively. Finally, the intersection of a z-simplex, s-simplex, and $\mathbb{Z}_+^3$ is our polytope $\mathbf{R}^{-1}((s, z)')$, the set of SFS that lie along the line $(x_1, z - 3s, -z + 4s - x_1)$. In Fig. 7A, as z ranges over $\{30, 31, \ldots, 40\}$, (1) the z-specific hue of the set of balls depicting the set $\mathbf{R}^{-1}((10, z)')$ ranges over $\{\text{violet}, \text{blue}, \ldots, \text{yellow}\}$, (2) $|\mathbf{R}^{-1}((10, z)')|$ ranges over $\{11, 10, \ldots, 1\}$ and (3) Tajima's d ranges over $\{-0.83, -0.53, \ldots, +2.22\}$, respectively. For example, there are eleven SFS in $\mathbf{R}^{-1}((10, 30)')$ and their Tajima's $d = -0.83$ (purple balls in Fig. 7A) and there is only one SFS in $\mathbf{R}^{-1}((10, 40)') = \{(0, 10, 0)\}$ such that its Tajima's $d = +2.22$ (yellow ball in Fig. 7A).

Analogously, when $n = 5$, we can project the first three coordinates of x, since $x_4 = s - x_1 - x_2 - x_3$. The intersection of the s-simplex, z-simplex and $\mathbb{Z}_+^4$ gives our set $\mathbf{R}^{-1}((s, z)')$ in the rectangular polytope via the parametric equation $(x_1, x_2, z/2 - 2s - x_2, 3s - z/2 - x_1)$ with $0 \le x_1 \le 3s - z/2$, $0 \le x_2 \le s$. In Fig. 7B, as z ranges over $\{20, 22, 24, 26, 28, 30\}$, (1) the z-specific hue of the set of balls depicting the set $\mathbf{R}^{-1}((5, z)')$ in the projected polytope ranges over $\{\text{violet}, \text{blue}, \ldots, \text{yellow}\}$, (2) $|\mathbf{R}^{-1}((5, z)')|$ ranges over $\{6, 10, 12, 12, 10, 6\}$ and (3) Tajima's d ranges over $\{-1.12, -0.56, 0.00, +0.56, +1.69\}$, respectively.

Unfortunately, $|\mathbf{R}^{-1}((s,z)')|$ grows exponentially with n and for any fixed n it grows geometrically with s. Thus, it becomes impractical to explicitly obtain $\mathbf{R}^{-1}(r)$ for reasonable sample sizes ($n > 10$). For small sample sizes, we used *Barvinok's cone decomposition* algorithm (Barvinok 1994) as implemented in the software package `LattE` (Loera et al. 2004) to obtain $|\mathbf{R}^{-1}((s,z)')|$ for 1000 data sets simulated under the standard neutral n-coalescent (Hudson 2002) with the scaled mutation rate $\phi_1^* = 10$. As n ranged in $\{4, 5, \ldots, 10\}$, the maximum of $|\mathbf{R}^{-1}((s,z)')|$ over the 1000 simulated data sets of sample size n ranged in:

$$\{73, 940, 6178, 333732, 1790354, 62103612, 190176626\},$$

respectively. Thus, even for samples of size 10, there can be more than 190 million SFS with exactly the same s and z. The SFS data in this simulation study with $\phi_1^* = 10$ corresponds to an admittedly long stretch of non-recombining DNA sites. On the basis of average per-site mutation rate in humans, this amounts to simulating human SFS data from n individuals at a non-recombining locus that is 100 kbp long, i.e., $m = 10^5$. Although such a large m is atypical for most non-recombining loci, it does provide a good upper bound for m and computational methods developed under a good upper bound are more likely to be efficient for smaller m. Our choices of ϕ_1^* and m are biologically motivated by a previous study on human SNP density (Sainudiin et al. 2007).

Thus, $|\mathbf{R}^{-1}((s,z)')|$ can make explicit computations over $\mathbf{R}^{-1}(r)$ impractical, especially for larger n. However, there are two facts in our favor: (1) if we are only interested in an expectation over $\mathbf{R}^{-1}(r)$ (with respect to some concentrated density) for reasonably sized samples (e.g. $4 \leq n \leq 120$), then we may use a Markov basis of $\mathbf{R}^{-1}(r)$ to facilitate Monte Carlo integration over $\mathbf{R}^{-1}(r)$ and (2) for specific summaries of SFS, such as the folded SFS $y := (y_1, y_2, \ldots, y_{\lfloor n/2 \rfloor})$, where $y_j := \mathbb{1}_{\{j \neq n-j\}}(j)\, x_j + x_{n-j}$, one can specify the Markov basis for any n.

The number of moves $|\mathcal{M}_{\mathbf{R}}|$ ranged over $\{2, 4, 6, 8, 14, 12, 26, 520, 10132\}$ as n ranged over $\{4, 5, \ldots, 9, 10, 30, 90\}$, respectively. The Markov basis for $\mathbf{R}^{-1}(r)$ when $n = 4$ is $\mathcal{M}_{\mathbf{R}} = \{(+1, 0, -1), (-1, 0, +1)\}$. From the example of Fig. 7A, we can see how $\mathbf{R}^{-1}(r)$ can be turned into a connected graph by $\mathcal{M}_{\mathbf{R}}$ for every r with $S = 10$. For instance, when $r = (10, 36)'$,

$$\mathbf{R}^{-1}(r) = \big\{(0, 6, 4), (1, 6, 3), (2, 6, 2), (3, 6, 1), (4, 6, 0)\big\}$$

and we can reach a neighboring SFS $\tilde{x} \in \mathbf{R}^{-1}(r)$ from any SFS $x \in \mathbf{R}^{-1}(r)$ by adding $(+1, 0, -1)$ or $(-1, 0, +1)$ to x, provided the sum is non-negative. When the sample size $n = 5$, a Markov basis for $\mathbf{R}^{-1}(r)$ is

$$\mathcal{M}_{\mathbf{R}} = \big\{(+1, 0, 0, -1), (-1, 0, 0, +1), (0, +1, -1, 0), (0, -1, +1, 0)\big\}$$

and once again we can see from Fig. 7B that any element $m \in \mathcal{M}_{\mathbf{R}}$ can be added to any $x \in \mathbf{R}^{-1}(r)$, for any r, to reach a neighbor within $\mathbf{R}^{-1}(r)$, *proviso quod*, $x_i + m_i \geq 0, \forall i$. Note that the maximum possible neighbors of any $x \in \mathbf{R}^{-1}(r)$ is bounded from above by $|\mathcal{M}_{\mathbf{R}}|$.

3.4.4 Folded Site Frequency Spectrum

The folded site frequency spectrum or FSFS $y := (y_1, y_2, \ldots, y_{\lfloor n/2 \rfloor})$ is essentially the SFS when one does not know the ancestral state of the nucleotide. It is determined by the map $Y(x) = y : \mathcal{X}_n^m \to \mathcal{Y}_n^m$:

$$\begin{aligned} Y(x) &:= \big(Y_1(x), Y_2(x), \ldots, Y_{\lfloor n/2 \rfloor}(x)\big), \\ Y_j(x) &:= x_j \mathbb{1}_{\{j \neq n-j\}}(j) + x_{n-j}, \ j \in \big\{1, 2, \ldots, \lfloor n/2 \rfloor\big\}. \end{aligned} \tag{33}$$

Y is the statistic of the n-coalescent experiment $\mathfrak{X}_{011} := (\mathcal{Y}_n^m, \sigma(\mathcal{Y}_n^m), \mathcal{P}_{\boldsymbol{\Phi}})$. The case of the FSFS y is particularly interesting since a Markov basis is known for any sample size n. Let e_i be the ith unit vector in $\mathbb{Z}^{n-1}$. A Markov basis of the set of y-preserving SFS $\mathbf{Y}^{-1}(y) := \{x : \mathbf{Y}x = y\}$ can be obtained by considering the null space of the matrix $\mathbf{Y}$, whose ith row $\mathbf{Y}_i$ is

$$\mathbf{Y}_i = \mathbb{1}_{\{j : j \neq (n-j)\}}(i)\, e_i + e_{n-i}, \quad i = 1, 2, \ldots, \lfloor n/2 \rfloor.$$

A minimal Markov basis $\mathcal{M}_{\mathbf{Y}}$ for $\mathbf{Y}^{-1}(y)$ is known explicitly for any n and contains the union of the following $2\lfloor n/2 \rfloor$ moves only:

$$\begin{cases} m_i = e_i - e_{n-i}, & i = 1, 2, \ldots, \lfloor n/2 \rfloor, \\ m_{n-i} = -e_i + e_{n-i}, & i = 1, 2, \ldots, \lfloor n/2 \rfloor. \end{cases}$$

The following algorithm can be used to make irreducible random walks in $\mathbf{Y}^{-1}(y)$: (i) Given an SFS x with folded SFS y, (ii) Uniformly pick $j \in \{1, 2, \ldots, \lfloor (n-1)/2 \rfloor\}$, (iii) Uniformly pick $k \in \{j, n-j\}$, (iv) Add $+1$ to x_k and add -1 to $x_{(n-k)}$, provided $x_{(n-k)} - 1 \geq 0$, to obtain an y-preserving SFS $\tilde{x}$ from x.

Note that x and $\tilde{x}$ have the same folded SFS y and fixing y also fixes s, z, Tajima's d and other summaries that are symmetric linear combinations of the SFS x. Thus, $\mathcal{M}_{\mathbf{Y}} \subseteq \mathcal{M}_{\mathbf{R}}$. For instance, when $n = 3$, $\mathcal{M}_{\mathbf{Y}} = \mathcal{M}_{\mathbf{R}} = \{(-1, +1), (+1, -1)\}$ and we have already seen that $\mathcal{M}_{\mathbf{Y}} = \mathcal{M}_{\mathbf{R}}$ when $n = 4, 5$. However, when $n \geq 6$ we may not necessarily have such an equality, i.e., $\mathcal{M}_{\mathbf{Y}} \subsetneq \mathcal{M}_{\mathbf{R}}$. When $n = 6$, our $\mathcal{M}_{\mathbf{R}}$ has extra moves so that $\mathcal{M}_{\mathbf{R}} \setminus \mathcal{M}_{\mathbf{Y}} = \{(+1, -4, +3, +0, +0), (-1, +4, -3, +0, +0)\}$. The size of the set $\mathbf{Y}^{-1}(y)$ follows from a basic permutation argument as

$$\big|\mathbf{Y}^{-1}(y)\big| = \prod_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} (y_i + 1).$$

3.4.5 Other Linear Experiments of the Site Frequency Spectrum

In principle, we can compute a Markov basis for any conditional lattice $\mathbf{G}^{-1}(g)$, such that $\mathbf{G}x = g \in \mathbb{Z}^k$, for some $k \times (n-1)$ matrix $\mathbf{G} := (g_{i,j})$, $g_{i,j} \in \mathbb{Z}_+$. Specifically, it is straightforward to add other popular summaries of the SFS. Examples of such linear summaries range from the unfolded singletons x_1, folded singletons $y_1 := x_1 + x_{(n-1)}$ (Hudson 1993) and Fay and Wu's $\theta_H := (n(n-1))^{-1} \sum_{i=1}^{n-1} (2\, i^2\, x_i)$ (Fay and Wu 2000).

3.4.6 Integrating over Neighborhoods of Site Frequency Spectra

Recall that a Markov basis $\mathcal{M}_{\mathbf{R}}$ for an observed linear summary r_o of the observed SFS x_o may be used to integrate some target distribution of interest over the set $\mathbf{R}^{-1}(r_o) := \{x \in \mathbb{Z}_+^{n-1} : \mathbf{R}x = r_o\}$. Such an integration may be conducted deterministically or stochastically. A simple deterministic strategy may entail a depth-first or a breadth-first search on the graph $\mathcal{G}_{\mathbf{R}}^{r_o}$ associated with the set $\mathbf{R}^{-1}(r_o)$ after initialization at x_o. A simple stochastic strategy may entail the use of moves in $\mathcal{M}_{\mathbf{R}}$ as local proposals for a Monte Carlo Markov chain sampler (MCMC) that is provably irreducible on $\mathbf{R}^{-1}(r_o)$. Such an MCMC sampler can be constructed, via the Metropolis–Hastings kernel for instance, to asymptotically target any distribution over the set $\mathbf{R}^{-1}(r_o) := \{x \in \mathbb{Z}_+^{n-1} : \mathbf{R}x = r_o\}$. Since every SFS state visited by such an MCMC sampler is guaranteed to exactly satisfy r_o, provided the algorithm is initialized at the observed SFS x_o and quickly converges to stationarity, one may hope to vanish the acceptance-radius ϵ altogether in practical approximate Bayesian computations that employ linear summaries of the SFS. One may use standard algebraic packages to compute $\mathcal{M}_{\mathbf{R}}$ for reasonably large sample sizes ($n < 200$). Furthermore, for perfectly symmetric summaries such as the folded SFS y we know a Markov basis for any n.

Unfortunately, the methodology is not immune to the curse of dimensionality. The set's cardinality ($|\mathbf{R}^{-1}((s,z)')|$) grows exponentially with n and for any fixed n it grows geometrically with the number of segregating sites s. This makes exhaustive integration of a target distribution over $\mathbf{R}^{-1}(r_o)$ impractical even for samples of size 10 with a large number of segregating sites. Also, even if we were to approximate the integral via Monte Carlo Markov chain with local proposals from the moves in $\mathcal{M}_{\mathbf{R}}$, the number of possible neighbors for some points in $\mathbf{R}^{-1}(r_o)$ may be as high as $|\mathcal{M}_{\mathbf{R}}|$. For instance, when the sample size $n = 90$, we may have up to 10, 132 moves. Such large degrees can lead to poor mixing of the MCMC sampler, especially when the initial condition is at the tail of the target distribution. However, there are some blessings that counter these curses. Firstly, the concentration of the target distribution under the n-coalescent greatly reduces the effective support on $\mathbf{R}^{-1}(r_o)$. Secondly, we can be formally interpolative in our integration strategy by exploiting the graph $\mathcal{G}_{\mathbf{R}}^{r_o}$ associated with the set $\mathbf{R}^{-1}(r_o)$ and the observed SFS x_o. Instead of integrating a target distribution over all of $\mathbf{R}^{-1}(r_o)$, either deterministically or stochastically, we can integrate over a ball of edge radius α about the observed SFS x_o:

$$\mathbf{R}_{\alpha}^{-1}(r_o) := \left\{x \in \mathbb{Z}_+^{n-1} : \mathbf{R}x = r_o, \|x - x_o\| \leq \alpha\right\},$$

where $\|x - x_o\|$ is the minimum number of edges between an SFS x and the observed SFS x_o. This integration over $\mathbf{R}_{\alpha}^{-1}(r_o)$ may be conducted deterministically via a simple breadth-first search on the graph $\mathcal{G}_{\mathbf{R}}^{r_o}$ associated with the set $\mathbf{R}^{-1}(r_o)$ by initializing at x_o. When a deterministic breadth-first search becomes inefficient, especially for large values of α, one may supplement with a Monte Carlo sampler that targets the distribution of interest over $\mathbf{R}_{\alpha}^{-1}(r_o)$. Since $\mathbf{R}_0^{-1}(r_o) = \{x_o\}$ and $\mathbf{R}_{\infty}^{-1}(r_o) = \mathbf{R}^{-1}(r_o)$, one can think of $\mathbf{R}_{\alpha}^{-1}(r_o)$ itself as an α-family of summary statistics that interpolates between the observed SFS x_o at one extreme and the observed coarser summary r_o

at the other. For a given observation x_o with its corresponding r_o and some reasonably large values of α, we can obtain $\mathbf{R}_\alpha^{-1}(r_o)$ independent of the target distribution via a single depth-first search. This is more efficient than a target-specific Monte Carlo integration over $\mathbf{R}_\alpha^{-1}(r_o)$ when we want to integrate multiple targets. Thus, we can integrate any target or set of targets over $\mathbf{R}_\alpha^{-1}(r_o)$ and thereby measure the extent of posterior concentration as α decreases from ∞ at one extreme to 0 at the other.

3.4.7 A Demographic Structured Population

Next, we demonstrate the generality of the methodology by studying a more complex model through linear summaries of more general summaries of the full data. For example, consider data from two known subpopulations A and B with sample sizes n^A and n^B, respectively, such that $n = n^A + n^B$. We can first summarize the data d_o into three vectors x^A, x^B and x^{AB} that can be thought of as a decomposition of the SFS based on subpopulations. Unlike the full SFS $x \in \mathbb{Z}_+^{n-1}$,

$$\begin{aligned} x^A &:= \left(x_1^A, \ldots, x_{n^A}^A\right) \in \mathbb{Z}_+^{n^A}, \\ x^B &:= \left(x_1^B, \ldots, x_{n^B}^B\right) \in \mathbb{Z}_+^{n^B}, \\ x^{AB} &:= \left(x_2^{AB}, \ldots, x_{n-1}^{AB}\right) \in \mathbb{Z}_+^{n-2}, \end{aligned}$$

where x_i^J is the number of sites that have i samples only from subpopulation $J \in \{A, B\}$ sharing a mutation (there are no mutations at these sites in the other subpopulation). We can think of x^A and x^B as subpopulation specific SFS and x^{AB} as the shared SFS. Thus, x_i^{AB} is the number of sites with a total of i samples (at least one sample from each population) having a mutation. Observe that the full SFS x for the entire sample can be recovered from the sub-population determined components as follows:

$$\begin{aligned} x_1 = x_1^A + x_1^B, \qquad x_2 = x_2^A + x_2^B + x_2^{AB}, \ldots, \qquad x_i = x_i^A + x_i^B + x_i^{AB}, \ldots, \\ x_{n-1} = x_{n-1}^{AB}. \end{aligned}$$

Now, let S^A, S^B, and S^{AB} be the number of segregating sites for A-specific, B-specific, and shared SFS, i.e.,

$$S^A := \sum_{i=1}^{n^A} x_i^A, \qquad S^B := \sum_{i=1}^{n^B} x_i^B, \quad \text{and} \quad S^{AB} := \sum_{i=2}^{n-1} x_i^{AB}.$$

Note that the total number of segregating sites is

$$S = \sum_{i=1}^{n-1} x_i = S^A + S^B + S^{AB}.$$

We are interested in the subpopulation determined SFS $\ddot{x}$ given by

$$\ddot{x} := \left(x^A, x^B, x^{AB}\right) = \left(x_1^A, \ldots, x_{n^A}^A, x_1^B, \ldots, x_{n^B}^B, x_2^{AB}, \ldots, x_{n-1}^{AB}\right) \in \mathbb{Z}_+^{2n-2}.$$

We refer to $\ddot{x}$ as the structured SFS (SSFS).

Let the non-averaged pair-wise heterozygosity be z for the entire sample and be z^A and z^B for sites segregating only within sub-population A and B, respectively, i.e.

$$z^A := \sum_{i=1}^{n^A-1} i\left(n^A - i\right)x_i^A, \quad \text{and} \quad z^B := \sum_{i=1}^{n^B-1} i\left(n^B - i\right)x_i^B.$$

Thus, the matrix $\mathbf{R}$ encoding the summary $r = (S^A, S^B, S^{AB}, z^A, z^B, z)$ is:

$$\mathbf{R} := \begin{pmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & \ldots & 0 & 1 & \ldots & 1 \\ 1(n^A-1) & \ldots & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1(n^B-1) & \ldots & 0 & 0 & \ldots & 0 \\ 1(n-1) & \ldots & n^A(n-n^A) & 1(n-1) & \ldots & n^B(n-n^B) & 2(n-2) & \ldots & n-1 \end{pmatrix}.$$

Observe that Tajima's D for the entire sample as well as the subpopulation specific D^A and D^B computed from the sites that are segregating only within subpopulation A and B, respectively, are also constrained by the six summaries. We could naturally add other linear summaries of x, x^A, x^B, and x^{AB}.

Finally, we can compute a Markov basis for $\mathbf{R}$ and use it to run Monte Carlo Markov chains on $\mathbf{R}^{-1}(r) = \{\ddot{x} : \mathbf{R}\ddot{x} = r\}$. The final ingredient we need is the target distribution on $\mathbf{R}^{-1}(r)$ when given some structured n-coalescent tree ${}^c\ddot{t}$ simulated according to ϕ, i.e., we need the probability $P(\ddot{x}|{}^c\ddot{t})$. This is also a Poisson multinomial distribution analogous to the simpler case with the sample SFS. However, the compression is not as simple as the total tree length ($l_\bullet$) and the relative time leading to singletons, doubletons, ..., "$n-1$-tons" ($\bar{l} \in \Delta_{n-2}$). Now, we need to divide the total length $l_\bullet$ of the tree ${}^c\ddot{t}$ into the length of lineages leading to mutations in subpopulation A alone ($l_\bullet^A$), in sub-population B alone ($l_\bullet^B$) and those leading to mutations in both subpopulations ($l_\bullet^{AB}$). Note that $l_\bullet = l_\bullet^A + l_\bullet^B + l_\bullet^{AB}$. The products of these three lengths $l_\bullet^A$, $l_\bullet^B$, and $l_\bullet^{AB}$ with ϕ_1 specifies the Poisson probability of observing S^A, S^B, and S^{AB}, respectively. To get the multinomial probabilities of x^A, x^B, and x^{AB}, we do a subpopulation-labeled compression of the structured n-coalescent tree ${}^c\ddot{t}$ into points in three simplexes. First, we label all the lineages of ${}^c\ddot{t}$ leading exclusively to mutations in subpopulation A. Next we compress these labeled lineages into the relative time leading to singletons, doubletons, ..., "n^A-tons" exclusively within subpopulation A. These labeled relative times yield $\bar{l}^A \in \Delta_{n^A-1}$. By an analogous labeling and compression of ${}^c\ddot{t}$ we obtain $\bar{l}^B \in \Delta_{n^B-1}$. Finally, we obtain the probabilities $\bar{l}^{AB} \in \Delta_{n-3}$ by labeling the lineages on ${}^c\ddot{t}$ that lead to both subpopulations.

3.5 n-Coalescent Experiments Graph

Having defined each one of the n-coalescent experiments, we next define a graph of n-coalescent experiments. This experiments graph sets a unified decision-theoretic stage that allows one to appreciate the different asymptotic senses and the partially

ordered graph of sub-σ-algebras or graph filtrations that underlie these classical experiments in population genetics.

Definition 2 (The Experiments Graph) Consider $\{\mathfrak{X}_\alpha, \alpha \in \mathfrak{A}\}$, an $\mathfrak{A}$-indexed set of experiments. Let, $T_{\alpha,\beta} : \mathcal{Z}_\alpha \to \mathcal{Z}_\beta$, 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 $T^{\circ}_{\alpha,\beta} := T_{\alpha,i} \circ T_{i,j} \circ \cdots \circ T_{i',j'} T_{j',\beta} : \mathcal{Z}_\alpha \to \mathcal{Z}_\beta$, such that $\sigma(\mathcal{Z}_\alpha) \supset \sigma(\mathcal{Z}_\beta)$. Then, by construction, (i) 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_α is $\sigma(\mathcal{X}_\alpha)$-measurable, such that (ii) $E(|X_\alpha|) < \infty$ for all $\alpha \in \mathfrak{A}$, form a martingale relative to $\mathcal{P}_{\boldsymbol{\Phi}}$ 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 Sects. 3.1, 3.2, 3.3, and 3.4, 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., $n \to \infty$. The second one involves the sequential limit in the number of sites $m \in \mathbb{N}$, i.e., $m \to \infty$. 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., $k \to \infty$. 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 $\to \infty$ and several bi/trisequential mixed senses of $(n, m, k) \to (\infty, \infty, \infty)$ with distinct asymptotic rates of convergence that are of decision-theoretic interest. See Felsenstein (2006) 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 $n \to \infty$, $m \to \infty$, and/or $k \to \infty$ 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 nonexistent over two decades ago, the large values of n, m, and k one encounters today were nonexistent half a decade ago and empirical resolutions that are much finer than our finest resolution of gap-free MSA are readily available today. However, population genomic inference at the finer resolutions of $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$, say at the currently realistic scale of one

thousand human genomes, is computationally prohibitive. Popular computational alternatives today include ABC and ALC methods that conduct heuristic inference at coarser empirical resolutions. We show that by an appropriate controlled lumped coalescent Markov chain we can indeed conduct exact inference at intermediate empirical resolutions of $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$, such as, experiments about the SFS.

The decision problem of computationally efficient and asymptotically consistent 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 we can obtain a *lumped n-coalescent graph* that underpins the unified multiresolution n-coalescent of Sainudiin and Stadler (2009). Through this lumped n-coalescent graph, the companion structure in the hidden space of our n-coalescent experiments graph $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$, it is also possible to take decisions that fully exploit the partially ordered filtrations that are indexed by sub-graphs of $\mathfrak{G}_{\mathfrak{A},\mathfrak{M}}$.

4 Applications

We next provide brief applications in testing and estimation under the simplest settings. These simplest models are already highly combinatorially structured and pose inferential challenges. Also, they are natural null models that form the basis for various classical tests in population genomics. In our applications, we are purposely using information from exactly one locus, as opposed to taking the product experiment over k loci that are assumed to have infinite recombination between them with zero intralocus recombination. The reason for our single locus design is to shed light on the algebraic statistical structure of the hidden space, particularly when it is ignored, during genome-scans for "unusual" loci. It is straightforward to extend our methods to k independent loci.

4.1 Topologically-conditioned Tests of Standard Neutrality

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, simulations may be used to calculate $P(D \leq d)$, where d is the observed value of D for a particular locus, under the standard neutral null model. 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 ${}^{c}t$ (Hudson 1993). In the empirical literature, there are a number of publications applying this procedure in order to discover "unusual" loci (reviewed in Thornton et al. 2007) that deviate from the null hypothesis of standard neutrality, i.e., a locus free of intra-locus recombination that is neutrally evolving in a large Wright–Fisher population of constant size under the IMS mutation model. Such topologically-unconditioned genome scans may be improved greatly at little

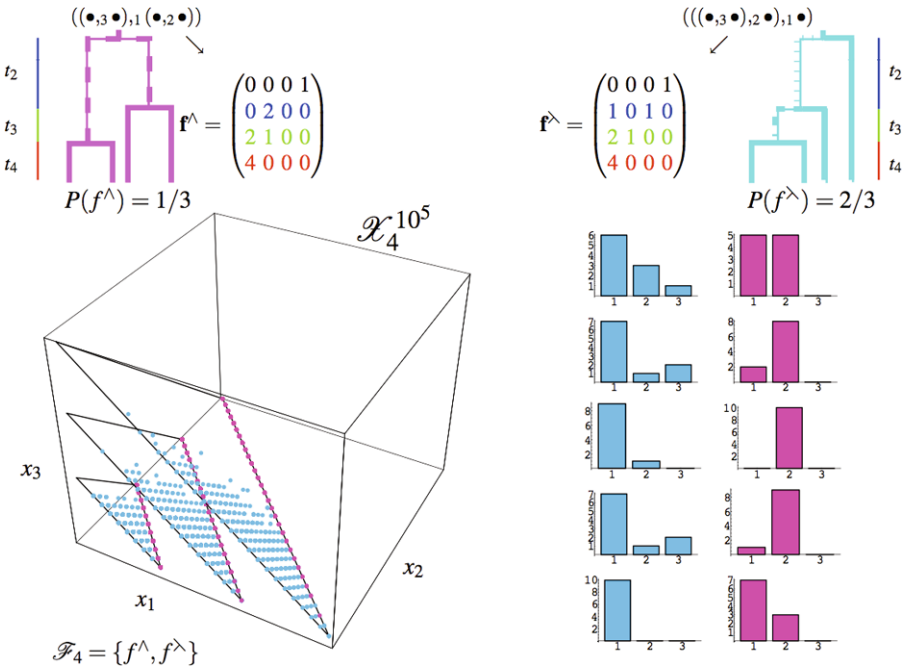

Fig. 8 Topological unfolding of SFS and Tajima's D. See text for description

additional computational cost. This can be achieved by conditioning on the partial topological information contained in $X^{\circledast}(x) = x^{\circledast}$ corresponding to the SFS x and employing Algorithm 2 to obtain topologically-conditioned null distributions of test statistics that are functions of the SFS.

Figure 8 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 $f^{\wedge\rangle}$ 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 (9) or (12). Thus, $P(c) = 1/18$ for each $c \in \mathcal{C}_n$ while $P(f^{\wedge\rangle}) = 2/3$ and $P(f^{\wedge}) = 1/3$. Five SFS simulations upon $f^{\wedge\rangle}$ and $f^{\wedge}$ are shown as the left and right columns of bar charts, respectively, on the lower right corner of Fig. 8. The remaining simulated SFS are plotted in the simplexes 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 $x_3 = 0$ and therefore $x_3^{\circledast} = 0$, as opposed to those SFS simulated under $f^{\wedge\rangle}$. Crucially, if we do not know the hidden $f \in \{f^{\wedge\rangle}, f^{\wedge}\}$ that the observed SFS x was realized upon, then the observation that $x_3 > 0$ implies that $x_3^{\circledast} = 1$, and this allows us to unambiguously eliminate $f^{\wedge}$ from the hidden space of f-sequences we need to integrate over or conditionally simulate from. This set of $x^{\circledast}$-specific hidden f-sequences is exactly $\complement\mathcal{F}_n(x^{\circledast})$ that we can access with the proposal Markov chain $\{F^{\lfloor x^{\circledast}}(k)\}_{k\in[n]_+}$ and its importance-reweighed

Table 2 10^4 loci were simulated under each hypothesised model $H_0, H_1, \ldots, H_8$ and tested for the extremeness of the observed Tajima's D statistic with and without conditioning on the observed $x^{\circledast}$ in an attempt to reject the null hypothesis H_0 at significance level $\alpha = 5\%$

Model: parameters	Proportion of loci rejected by null distribution of test statistics			
$H_i : (m\phi_1, \phi_2, \rho)$	$P_{H_0}(D \geq d)$	$P_{H_0}(D \geq d \mid x^{\circledast})$	$P_{H_0}(D \leq d)$	$P_{H_0}(D \leq d \mid x^{\circledast})$
$H_0 : (100, 0, 0)$	0.0495	0.0501	0.0499	0.0501
$H_1 : (100, 0, 10)$	0.0074	0.8640	0.0061	0.0017
$H_2 : (100, 0, 100)$	0.0000	0.9999	0.0000	0.0000
$H_3 : (100, 10, 0)$	0.0000	0.0019	0.0326	0.1759
$H_4 : (100, 10, 10)$	0.0001	0.2023	0.0135	0.0797
$H_5 : (100, 10, 100)$	0.0000	0.5559	0.0006	0.0180
$H_6 : (100, 100, 0)$	0.0000	0.0000	0.1696	0.6882
$H_7 : (100, 100, 10)$	0.0000	0.0002	0.1580	0.6668
$H_8 : (100, 100, 100)$	0.0000	0.0020	0.1321	0.6617

variants. Thus, by means of Algorithm 2 that invokes $\{F^{\downarrow x^{\circledast}}(k)\}_{k \in [n]_+}$ and further reweighing by $P(f)$ we can generate the topologically conditioned null distribution of any statistic that is a function of SFS, including the classical linear combinations of Sect. 3.4 as well as various classical and non-classical tree shape statistics (Sainudiin and Stadler 2009).

The power of classical Tajima's D test with that of its topologically conditioned version is compared in Table 2. The significance level α is set at 5% for the standard neutral null hypothesis H_0 and eight alternative hypotheses, namely, $H_1, \ldots, H_8$, were explored by increasing the recombination rate and/or the growth rate with parameters as shown in Table 2. Here, $m\phi_1$ is the scaled per-locus mutation rate, ϕ_2 is the exponential growth rate and ρ is the scaled per-locus recombination rate. The $x^{\circledast}$-conditional tests based on Tajima's D are more powerful than the unconditional classical tests since a larger proportion of the 10^4 loci simulated under the alternative models are rejected. All simulations were conducted using standard coalescent methods (Hudson 2002).

4.2 Exactly Approximate Likelihoods and Posteriors

In computational population genetics, an approximate likelihood or an approximate posterior merely refers to the exact likelihood or the exact posterior based on some statistic $R(v) = r : \mathcal{V}_n^m \to \mathcal{R}_n^m$. R is called a *summary statistic* to emphasize the fact that it may not be sufficient. Approximating the likelihood of the observed statistic r_o is often a computationally feasible alternative to evaluating the likelihood of the observed data v_o. Here, *approximate* is meant in the hopeful sense that R may not be a sufficient statistic, i.e., in the Bayesian sense that $P(\phi|v) \neq P(\phi|r = R(v))$, but perhaps approximately sufficient, i.e., $P(\phi|v) \approxeq P(\phi|r)$ under some reasonable criterion. The exact evaluation of the approximate posterior $P(\phi|r)$ involves the exact evaluation of the likelihood $P(r|\phi)$ with standard errors. For an arbitrary statistic R, such exact evaluations may not be trivial. However, one may resort to the follow-

ing simulation-based inferential methods termed approximate Bayesian or likelihood computations in order to approximately evaluate $P(\phi|r)$ or $P(r|\phi)$, respectively.

4.2.1 ABC

In approximate Bayesian computation or ABC (Beaumont et al. 2002), one typically simulates data $v \in \mathcal{V}_n^m$ with a ϕ-indexed family of measures, such as the Kingman's n-coalescent superimposed by Watterson's infinitely-many-sites mutations, after drawing a ϕ according to its prior distribution $P(\phi)$, then summarizes it to $r = R(v) \in \mathcal{R}_n^m$ and finally accepts ϕ if $\mathfrak{m}(r, r_o) \leq \epsilon$, where the map $\mathfrak{m} : \mathcal{R}_n^m \times \mathcal{R}_n^m \to \mathbb{R}_+$ is usually a metric on $\mathcal{R}_n^m$ and ϵ is some nonnegative acceptance-radius. Algorithm 3 details one of the simplest ABC schemes. Approximate likelihood computation or ALC (Weiss and von Haeseler 1998) is similar to ABC, except one typically conducts the simulations over a finite uniform grid of G points in the parameter space $\boldsymbol{\Phi}$ denoted by $\boldsymbol{\Phi}_G = \{\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(G)}\}$. In a simple ALC, one distributes the computational resources evenly over the G parameters in $\boldsymbol{\Phi}_G$ and approximates the likelihood at $\phi^{(i)}$ by the proportion of times the summary r of a data v simulated under $\phi^{(i)}$ was accepted on the basis of $\mathfrak{m}(r, r_o) \leq \varepsilon$. As the grid size and the number of simulations increase, the likelihood estimates based on ALC are indistinguishable from the posterior estimate based on ABC under a uniform prior on the appropriate hyper-cuboid containing $\boldsymbol{\Phi}_G$.

Algorithm 3 A simple ABC/ALC algorithm

input:

1. a samplable distribution $P(v|\phi)$ over $\mathcal{V}_n^m$ indexed by $\phi \in \boldsymbol{\Phi}$
2. a samplable prior $P(\phi)$
3. observed data $v_o \in \mathcal{V}(v)_n^m$ and summaries $r_o = R(v_o) \in \mathcal{R}_n^m$
4. tolerance $\varepsilon \geq 0$
5. a map $\mathfrak{m} : \mathcal{R}_n^m \times \mathcal{R}_n^m \to \mathbb{R}_+$
6. a large positive integer $\texttt{MAXTRIALS} \in \mathbb{N}$

output: a sample $U \sim P(\phi|\mathbf{r}_\varepsilon(r_o)) \approxeq P(\phi|r_o) \approxeq P(\phi|v_o)$ or {}, where, $\mathbf{r}_\varepsilon(r_o) := \{r : \mathfrak{m}(r, r_o) \leq \varepsilon\}$.
initialize: $\texttt{TRIALS} \leftarrow 0$, $\texttt{SUCCESS} \leftarrow \texttt{false}$, $U \leftarrow \{\}$
repeat
  $\phi \leftarrow P(\phi)$ {DRAW from Prior}
  $v \leftarrow P(v|\phi)$ {SIMULATE data}
  $r \leftarrow R(v)$ {SUMMARIZE data}
  if $\mathfrak{m}(r, r_o) \leq \varepsilon$ **then** {COMPARE summaries and ACCEPT/REJECT parameter}
    $U \leftarrow \phi$, $\texttt{SUCCESS} \leftarrow \texttt{true}$
  end if
  $\texttt{TRIALS} \leftarrow \texttt{TRIALS} + 1$
until $\texttt{TRIALS} \geq \texttt{MAXTRIALS}$ or $\texttt{SUCCESS} \leftarrow \texttt{true}$
return: U

Statistical justification of ALC and ABC methods rely on the summary statistic R being close to the typically unknown sufficient statistic and thereby producing reasonably approximate likelihood and posterior. However, as R gets closer to the sufficient statistic one has to make the acceptance-radius ε unreasonably large to increase the acceptance rate of the proposed ϕ. For instance, current ALC and ABC methods have unacceptably low acceptance rates for a reasonably small ε if r is taken as the SFS. But when ε is too large we gain little information from the simulations.

Let us examine the "ε-dilemma" under the ABC framework in detail. Analogous arguments also apply for the ALC framework. In ABC, samples are drawn from an ε-specific approximation of $P(\phi|r_o)$. Since $\mathbf{r}_\varepsilon(r_o) := \{r : \mathfrak{m}(r, r_o) \leq \varepsilon\}$, we are making the following posterior approximation of the ultimately desired $P(\phi|v_o)$:

$$P(\phi|v_o) \approxeq \begin{cases} P(\phi|r_o) = P(\phi|\{v : R(v) = R(v_o) = r_o\}) & \text{if: } \varepsilon = 0, \\ P(\phi|\mathbf{r}_\varepsilon(r_o)) = P(\phi|\{v : \mathfrak{m}(R(v), R(v_o)) \leq \varepsilon\}) & \text{if: } \varepsilon > 0. \end{cases}$$

The assumed approximate sufficiency of the statistic R, i.e., $P(\phi|v_o) \approxeq P(\phi|r_o)$, terms the posterior $P(\phi|r_o)$ *approximate*. Furthermore, the nonzero acceptance-radius ε, for reasons of computational efficiency, yields the *further ε-specific approximate* posterior $P(\phi|\mathbf{r}_\varepsilon(r_o))$. In the extremal case, the approximate posterior $P(\phi|\mathbf{r}_\infty(r_o))$ equals the prior $P(\phi)$, and we have gained no information from the experiment. Furthermore, there is no guarantee that a computationally desirable metric $\mathfrak{m}$ is also statistically desirable, i.e., produce reasonably approximate posterior samples.

Considerable effort is expended in fighting this "ε-dilemma" by say (1) smoothing the $\mathfrak{m}(r, r_o)$'s (Beaumont et al. 2002) or (2) making use of local Monte Carlo samplers (Marjoram et al. 2003) or (3) finding the right sequence of ε's under the appropriate metric $\mathfrak{m}$ (Sisson et al. 2007) in order to obtain the optimal trade-off between efficiency and accuracy (see Bertorelle et al. 2010 for a recent review of ABC methods). It is difficult to ensure that such sophisticated battles against the "ε-dilemma" that arise in the simulation-based inferential approaches of ABC and ALC do not confound the true posterior $P(\phi|r_o)$ or the true likelihood $P(r_o|\phi)$. Thus, both ABC and ALC methods may benefit from exact methods that can directly produce the likelihood $P(r_o|\phi)$, for at least a class of summary statistics. They may also benefit from a systematic treatment of the relative information in different sets of summary statistics obtainable with exact methods.

4.2.2 ABCDE

For a large class of statistics, namely the SFS and its various linear combinations, our approach allows the acceptance radius ε to equal zero. This is achieved by Monte Carlo simulations of the controlled lumped coalescent Markov chain $\{F^{\lfloor x^{\circledast}}(k)\}_{k\in[n]_+}$ of Algorithm 2 and further reweighing by $P(f)$ to evaluate $P(x|\phi)$ in (25) and $P(r|\phi)$ in (29). Therefore, our approach yields an exact evaluation of the desired approximate posterior $P(\phi|r)$ and amounts to ABCDE or ABC done exactly.

Table 3 Performance of our estimator of $m\phi_1^*$ and ϕ_2^* based on SFS (see text)

n	Performance of $\widehat{m\phi_1}$			Performance of $\widehat{\phi_2}$			Performance of $(\widehat{m\phi_1}, \widehat{\phi_2})$	
	$\sqrt{\widehat{se}}$	bs	$C_{99\%}$	$\sqrt{\widehat{se}}$	bs	$C_{99\%}$	$C_{99\%}$	Quartiles of $\breve{\mathcal{K}}$
4	40	28	.545	41	26	.185	.828	{0.062, 0.085, 0.143}
5	35	22	.584	34	20	.236	.832	{0.073, 0.102, 0.167}
6	30	19	.602	32	18	.343	.824	{0.081, 0.109, 0.178}
7	27	16	.660	29	14	.410	.838	{0.089, 0.126, 0.209}
8	23	13	.687	25	11	.474	.852	{0.096, 0.142, 0.235}
9	20	11	.712	23	10	.554	.872	{0.102, 0.155, 0.263}
10	19	10	.711	25	11	.604	.858	{0.106, 0.164, 0.294}

4.2.3 Parameter Estimation in an Exponentially Growing Population

We estimate the locus-specific scaled mutation rate $m\phi_1^*$ and the exponential growth rate ϕ_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 the standard neutral model with $m\phi_1^* = 10.0$ and $\phi_2^* = 0.0$ (for human data $m\phi_1^* = 10.0$ implies a locus of length 10 kbp, i.e., $m = 10^4$) (Hudson 2002). Our choices of ϕ_1^* and m are biologically motivated by a previous study on human SNP density (Sainudiin et al. 2007). Our point estimate $(\widehat{m\phi_1}, \widehat{\phi_2})$ of $(m\phi_1^*, \phi_2^*)$ 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 101×101 parameter points $\phi = (\phi_1, \phi_2)$ over our rectangular uniform prior density $((100 - 1/10000)100)^{-1} \mathbb{1}_{\{[0.0001,100],[0,100]\}}(\phi_1, \phi_2)$.

Our performance measures can help make natural connections to the theory of approximate sufficiency (Cam 1964), as we not only measure the bias (bs), root-mean-squared-error ($\sqrt{\widehat{se}}$) 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 $\breve{\mathcal{K}}$, the Kullback–Leibler divergence between the posterior histogram estimate and the uniform prior that is rescaled by the prior's entropy. Table 3 gives the maximum a posteriori estimate of $(m\phi_1^*, \phi_2^*)$ by a Monte Carlo sum over ϕ_2-specific epoch-time vectors in $\mathbb{T}_n := (0, \infty)^{n-1}$ and every $x^{\circledast}$-specific hidden f-sequence in $\mathbb{C}_{F_n}(x^{\circledast})$ by means of Algorithm 2 that invokes $\{F^{\downarrow x^{\circledast}}(k)\}_{k \in [n]_+}$.

We also obtained maximum a posteriori point estimate $(\widehat{m\phi_1}, \widehat{\phi_2})$ of $(m\phi_1^*, \phi_2^*) = (10, 0)$ based on (s, z) and $(s, z, x^{\circledast})$ of the SFS x. Our ABCDE estimators are equivalent to exactly approximate Bayesian computations (with $\varepsilon = 0$) as we integrate exhaustively over all SFS in $\mathbf{R}^{-1}((s, z)')$ when we compute $P(\phi|(s, z))$ or $P(\phi|(s, z, x^{\circledast}))$. For the same set of simulated data of Table 3 the joint empirical coverage significantly suffered at about 50% for the estimator that only used (s, z). By using additional topological information, the estimator based on $(s, z, x^{\circledast})$ had a better coverage that improved with sample size (between 61% and 76%). We also restrict the sample size to exhaustively integrate over the fiber $\mathbf{R}^{-1}(s, z')$ and avoid expositions on Monte Carlo samplers over $\mathbf{R}^{-1}(s, z')$ for brevity. Contrastingly, the

coverage was nearly perfect when the entire BIM was used to estimate the parameters through an importance sampler (Stephens and Donnelly 2000).

When ABC is done exactly, it is clear that using a few coarse linear summaries of the SFS, even after a topological conditioning by $x^{\circledast}$, is not only computationally inefficient but also provides significantly less information when compared to using the entire SFS. Nonetheless, these computations over population genetic fibers shed algebraic insights and provide exact benchmarks against which one can compare, correct and improve simulation-intensive ABC/ALC algorithms in the current molecular population genetic literature that ignore topological information up to sufficient equivalence classes in the hidden space of genealogies.

Author's Contributions KT and RS posed the ABCDE problem. RS developed the controlled lumped chain $\{F^{\lfloor x^{\circledast}}(k)\}_{k\in[n]_+}$ based on discussions with PD, RG, and GM. RS coded all related modules of `LCE-0.1` with support from KT. JH efficiently reimplemented Algorithm 2. RS developed the samplers over population genetic fibers based on discussions with JB, MS, and RY. RY and MS conducted the algebraic statistical computations in `LattE` and `Macaulay 2`, respectively. RS wrote the first draft and incorporated comments made by the coauthors.

Acknowledgements R.S. was supported by an NSF/NIGMS grant DMS-02-01037 and a research fellowship from the Royal Commission for the Exhibition of 1851 under the sponsorship of P.D. during the course of this study. J.H. was supported by an Allan Wilson Summer Studentship. R.Y. was supported by NIGMS grant 1R01GM086888. Comments from two anonymous referees greatly improved the manuscript. R.S. thanks Celine Becquet for discussions on summary statistics of structured populations, Mike Steel for (Kemeny 1960, Definition 6.3.1), Jesse Taylor for (Kingman 1982b, 5.2), Joe Watkins for the articulation of Definition 2, Michael Nussbaum and Simon Tavaré for discussions on approximate sufficiency, and Scott Williamson for generous introductions to the site frequency spectrum.

References

Bahlo, M., & Griffiths, R. (1996). Inference from gene trees in a subdivided population. *Theor. Popul. Biol. 57*, 79–95.

Barvinok, A. (1994). Polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. *Math. Oper. Res. 19*, 769–779.

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

Bertorelle, G., Benazzo, A., & Mona, S. (2010). ABC as a ßexible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol. 19*, 2609–2625.

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

Cam, L.L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Stat. 35*, 1419–1455.

Casanellas, M., Garcia, L., & Sullivant, S. (2005). Catalog of small trees. In L. Pachter & B. Sturmfels (Eds.), *Algebraic statistics for computational biology* (pp. 291–304). Cambridge: Cambridge University Press.

Diaconis, P., & Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Stat. 26*, 363–397.

Duflo, M. (1997). *Random iterative models*. Berlin: Springer.

Erdös, P., Guy, R., & Moon, J. (1975) On refining partitions. *J. Lond. Math. Soc. (2) 9*, 565–570.
Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol. 3*, 87–112.
Ewens, W. (1974). A note on the sampling theory of infinite alleles and infinite sites models. *Theor. Popul. Biol. 6*, 143–148.
Ewens, W. (2000). *Mathematical population genetics* (2nd edn.). Berlin: Springer.
Fay, J., & Wu, C. (2000). Hitchhiking under positive Darwinian selection. *Genetics 155*, 1405–1413.
Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol. 17*, 368–376.
Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol. Biol. Evol. 23*, 691–700.
Grayson, D., & Stillman, M. (2004). Macaulay 2, a software system for research in algebraic geometry. Available at www.math.uiuc.edu/Macaulay2.
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.
Griffiths, R., & Tavare, S. (2003). The genealogy of a neutral mutation. In P. Green, N. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 393–412). London: Oxford University Press.
Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.
Hemmecke, R., Hemmecke, R., & Malkin, P. (2005). 4ti2 version 1.2—computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. Available at www.4ti2.de.
Hosten, S., Khetan, A., & Sturmfels, B. (2005). Solving the likelihood equations. *Found Comput. Math. 5*(4), 389–407.
Hudson, R. (1993). The how and why of generating gene genealogies. In: Clark. A., Takahata, N. (Eds.) *Mechanisms of molecular evolution* (pp. 23–36). Sunderland: Sinauer.
Hudson, R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics 18*, 337–338.
Iorio, M., & Griffiths, R. (2004). Importance sampling on coalescent histories. I. *Adv. Appl. Probab.*, *36*, 417–433.
Jones, G., & Hobert, J. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Stat. Sci. 16*(4), 312–334.
Jukes, T., & Cantor, C. (1969). Evolution of protein molecules. In H. Munro (Ed.), *Mammalian protein metabolism* (pp. 21–32). San Diego: Academic Press.
Kemeny, Snell (1960). *Finite Markov chains*. Princeton: Van Nostrand.
Kendall, D. (1975). Some problems in mathematical genealogy. In: Gani, J. (Ed.), *Perspectives in probability and statistics* (pp. 325–345). San Diego: Academic Press.
Kingman, J. (1982a). The coalescent. *Stoch. Process. Their Appl. 13*, 235–248.
Kingman, J. (1982b). On the genealogy of large populations. *J. Appl. Probab. 19*, 27–43.
Kolmogorov, A. (1942). Sur l'estimation statistique des parametères de la loi de gauss. *Bull. Acad. Sci. URSS Ser. Math. 6*, 3–32.
Loera, J. D., Haws, D., Hemmecke, R., Huggins, P., Tauzer, J., & Yoshida, R. (2004). Lattice Point Enumeration: LattE, software to count the number of lattice points inside a rational convex polytope via Barvinok's cone decomposition. Available at www.math.ucdavis.edu/~latte.
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.
Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys. 21*, 1087–1092.
Mossel, E., & Vigoda, E. (2005). Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*, *309*, 2207–2209.
Mossel, E., & Vigoda, E. (2006). Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Ann. Appl. Probab.*, *16*(4), 2215–2234.
Rosenblatt, M. (1974). *Random processes*. Berlin: Springer.
Sainudiin, R., & Stadler, T. (2009) A unified multi-resolution coalescent: Markov lumpings of the Kingman-Tajima n-coalescent. UCDMS Research Report 2009/4, 5 April 2009 (submitted). Available at http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/SixCoal.pdf.
Sainudiin, R., & York, T. (2009). Auto-validating von Neumann rejection sampling from small phylogenetic tree spaces. *Algorithms Mol. Biol. 4*, 1.
Sainudiin, R., Clark, A., & Durrett, R. (2007). Simple models of genomic variation in human SNP density. *BMC Genomics 8*, 146.

Semple, C., & Steel, M. (2003). *Phylogenetics*. Oxford University Press, London.
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. Popul. Biol. 62*, 339–348.
Stephens, M., & Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Stat. Soc. B 62*, 605–655.
Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics 123*, 585–595.
Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol. 26*, 119–164.
Thornton, K., Jensen, J. D., Becquet, C., & Andolfatto, P. (2007). Progress and prospects in mapping recent selection in the genome. *Heredity 98*, 340–348.
Wakeley, J. (2007). *Coalescent theory: an introduction*. Greenwood Village: Roberts & Co.
Watterson, G. (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.
Yang, Z. (2000). Complexity of the simplest phylogenetic estimation problem. *Proc. R. Soc. Lond. B Biol. Sci. 267*, 109–119.