Coalescent experiments II: Markov bases of classical population genetic statistics

Raazesh Sainudiin[1], Kevin Thornton[2], James Booth[3], Michael Stillman[4] and Ruriko Yoshida[5]

[1] Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, NZ
[2] Department of Ecology and Evolutionary Biology, University of California, Irvine, USA
[3] Department of Biological Statistics and Computational Biology
[4] Department of Mathematics, Cornell University, Ithaca, USA
[5] Department of Statistics, University of Kentucky, Lexington, USA

UCDMS 2009/8 (May 19, 2009).

Abstract. Evaluating the likelihood function of parameters in complex population genetic models from extant deoxyribonucleic acid (DNA) sequences is computationally prohibitive. In such cases, one may approximately infer the parameters from various summary statistics of the data. Such method are known as approximate likelihood/Bayesian computations. We employ computational commutative algebraic methods to obtain the exact likelihood of a large class of summary statistics that are linear combinations of the site frequency spectrum.

keywords. integrating controlled coalescent measures via Markov bases; exactly approximate Bayesian/likelihood computation in population genetics

1 Introduction

This is a companion article to the prerequisite article [1]. Here we are interested in the posterior distribution over the parameter space Φ on the basis of various classical molecular population genetic statistics of DNA sequences obtained from a sample of n individuals in a population at m homologous sites. The statistics of interest at the coarsest end, include (1) the non-negative integer-valued *number of segregating sites* S [2], (2) the rational-valued *average heterozygosity* Π, (2') the real-valued Tajima's D [3] that combines (1) and (2). At a slightly finer resolution than the first three that is of interest is (3) the nonnegative integer

vector called the *folded site frequency spectrum* $Y \in \mathbb{Z}_+^{\lfloor n/2 \rfloor}$. These statistics as well as others, including folded singletons Y_1 [4] and Fay and Wu's Θ_H [5], are linear combinations [6] of the nonnegative integer vector called the *site frequency spectrum* or SFS X. We will exploit this linear relationship between SFS and the above statistics via Markov bases [7] in the context of evaluating their likelihoods. We can obtain the site frequency spectrum x of a given DNA sequence data u from a standard encoding of u into a binary incidence matrix or BIM v as described in [1, §2]. Recall that Kingman's unlabeled n-coalescent [1, (24)] is the basic probability model for $P(x|\phi)$.

The rest of the article is organized as follows. The basic form of the data, statistics, the underlying probability models and the associated inference methods are introduced in the prerequisite article [1] that the reader is expected to have read for background. The methodology for posterior inference from site frequency spectrum statistics via their Markov bases [7] is summarized in §2. We apply the methodology to linear statistics of the site frequency spectrum in §3 and present some results on simulated data in §3.5. After suggesting some natural extensions of inferential methods using Markov bases in §4 we conclude with a discussion in §5. A beta version of `LCE-0.1: A C++ class library for lumped coalescent experiments` that implements some algorithms described in this article is available from `http://www.math.canterbury.ac.nz/~r.sainudiin/codes/lce/` under the terms of the GNU General Public License.

2 Exactly Approximate Likelihoods and Posteriors

Computationally-intensive inference, based on an observed multiply aligned homologous DNA sequences $u_o \in \mathcal{U}_n^m := \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^{n \times m}$ from n individuals at m sites, with realistically large n and m, is currently infeasible for recombining loci and prohibitive for non-recombining loci. An alternative inference strategy that is computationally feasible involves a summary or a relatively low-dimensional statistic of the observed data $u_o \in \mathcal{U}_n^m$. In this approach, one attempts to approximate the likelihood $P(u_o|\phi)$ or the posterior distribution $P(\phi|u_o)$, on the basis of a summary $r_o = R(u_o)$ of the data u_o, where $R(u) = r : \mathcal{U}_n^m \mapsto \mathcal{R}_n^m$ is a statistic with $\mathcal{R}_n^m$ as its sample space. Since R is usually not a sufficient statistic for ϕ, i.e., $P(\phi|r_o) \neq P(\phi|u_o)$, such methods have been termed as approximate likelihood computations (ALC) [8] in a frequentist setting or as approximate Bayesian computations (ABC) [9, 10] 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 known to be sufficient or even necessarily consistent, form the basis of information in such approximate likelihood or Bayesian computations. The underlying assumption is that a large enough set of such statistics will be a good proxy for the observed data u_o, in the approximately sufficient sense see [1, §3] and [11]. Any formal notion of approximate sufficiency for population genetic data $u_o \in \mathcal{U}_n^m$ must account for the fact that the like-

lihood $P(u_o|\phi) = \sum \int_{^c t \in {}^{\mathcal{C}_n}\mathbb{T}_n} P(u_o|{}^c t, \phi) P({}^c t|\phi)$ is defined as the n-coalescent prior mixture over elements in a partially observed genealogical space ${}^{\mathcal{C}_n}\mathbb{T}_n$ of labeled n-coalescent trees [1, §3.2 and Fig. 3]. This space is both discrete, to account for the sequence of coalescence events, and continuous, to account for the number of generations between such events in units of rescaled time.

2.1 Approximate Likelihoods and Approximate 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 $R(v) = r : \mathcal{V}_n^m \to \mathcal{R}_n^m$ that summarizes the finer resolution of data $v \in \mathcal{V}_n^m$ available to us into $\mathcal{R}_n^m$, the sample space of the statistic R. Here, $\mathcal{V}_n^m$ is the space of binary incidence matrices (BIM's) and *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., $P(\phi|v) \approxeq P(\phi|r = R(v))$ under some reasonable criterion. The exact evaluation of the approximate posterior $P(\phi|r = R(v))$ involves the exact evaluation of the likelihood $P(r = R(v)|\phi)$ with standard errors. For an arbitrary statistic R, such exact evaluations may not be trivial. However, one may resort to the following simulation-based inferential methods termed approximate Bayesian or likelihood computations in order to approximately evaluate $P(\phi|r_o = R(v_o))$ or $P(r_o = R(v_o)|\phi)$, respectively, based on the observed statistic $r_o = R(v_o) \in \mathcal{R}_n^m$ that summarizes the observed data $v_o \in \mathcal{V}_n^m$.

In approximate Bayesian computation or ABC, one typically (1) simulates data $v \in \mathcal{V}_n^m$ with a ϕ-indexed family of measures, such as the Kingman's n-coalescent, after drawing a ϕ according to its prior distribution $P(\phi)$, (2) summarizes it to $r = R(v) \in \mathcal{R}_n^m$ and (3) 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 non-negative acceptance-radius. Algorithm 1 details one of the simplest ABC schemes. Approximate likelihood computation or ALC is similar to ABC, except one typically conducts the simulations over a finite uniform grid of G points in the parameter space $\mathbf{\Phi}$ denoted by $\mathbf{\Phi}_G = \{\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(G)}\}$. In a simple ALC, one distributes the computational resources evenly over the G parameters in $\mathbf{\Phi}_G$ 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 $\mathfrak{m}(r, r_o) \leq \epsilon$. 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 compact box containing $\mathbf{\Phi}_G$. In fact, we use such a prior in the sequel to address elementary statistical and computational issues shared by both frequentist and Bayesian paradigms within such simulation-based inferential methods .

In both ABC and ALC, as the summaries get finer (closer to the data) one has to make the acceptance-radius ϵ large to increase the acceptance rate of the proposed ϕ. However, when ϵ is too large we gain little information from the simulations. Considerable effort is expended in fighting this 'ϵ-dilemma' by say

Algorithm 1 A Simple ABC/ALC Algorithm

input:

1. a samplable distribution $P(v|\phi)$ over $\mathcal{V}_n^m$ indexed by $\phi \in \Phi$
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 $\epsilon \geq 0$
5. a map $\mathfrak{m} : \mathcal{R}_n^m \otimes \mathcal{R}_n^m \mapsto \mathbb{R}_+$
6. a large positive integer $\texttt{MAXTRIALS} \in \mathbb{N}$

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

(1) smoothing the $\mathfrak{m}(r, r_o)$'s [9] or (2) finding the right sequence of ϵ's under the appropriate metric $\mathfrak{m}$ in order to obtain the optimal trade-off between efficiency and accuracy. 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.

Next we examine the 'ϵ-dilemma' under the ABC framework more closely. Analogous arguments also apply for the ALC framework. In ABC, samples are drawn from an ϵ-specific approximation of $P(\phi|r_o)$. Since, $\mathbf{r}_\epsilon(r_o) := \{r : \mathfrak{m}(r, r_o) \leq \epsilon\} := \{v : \mathfrak{m}(R(v), R(v_o)) \leq \epsilon\}$, we are making the following posterior approximation of the ultimately desired $P(\phi|v_o)$:

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

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 non-zero acceptance-radius ϵ, for reasons of computational efficiency, yields the *further ϵ-specific approximate* posterior $P(\phi|\mathbf{r}_\epsilon(r_o))$. 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. In the sequel, we show that for any subset of statistics from a large class of statistics $\mathfrak{R}$, the acceptance-radius ϵ can be made to vanish altogether under any metric and we can exactly evaluate the likelihood $P(R_1(v) = r_1, \ldots, R_k(v) = r_k|\phi)$, where $R_i \in \mathfrak{R}$ for all

$i = 1, \ldots, k$. This yields an exact evaluation of the desired approximate posterior $P(\phi|R_1(v) = r_1, \ldots, R_k(v) = r_k)$.

Briefly, this is accomplished by restricting the class of summaries in $\mathfrak{R}$ to the site frequency spectrum and its linear combinations and employing a computational commutative algebraic approach involving Markov Bases [7] to facilitate integrations over all site frequency spectra that exactly satisfy a given set of observed statistics from $\mathfrak{R}$. In order to integrate over the appropriately sufficient equivalence classes in the genealogical space $\mathcal{C}_n$ of c-sequences, we employ the *unlabeled n-coalescent*, a continuous time Markov chain on the set of integer partitions of the sample size n [1, §3.3] and the associated Markov processes [1, §3.5]. These processes are sufficient and necessary to prescribe the needed family of measures on the sample space of the site frequency spectra [1, Prop. 8]. The reader is assumed to be familiar with [1, §3.3–3.5] before continuing to the next section.

2.2 Exact Likelihood and Exact Posteriors

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

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

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

We want to compute $P(r|\phi)$, since the posterior distribution of interest $P(\phi|r) \propto P(r|\phi)P(\phi)$. 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:

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

$$P(r|\phi) = \int_{l \in \mathcal{L}_n} P(r|\phi, l) P(l|\phi) = \int_{l \in \mathcal{L}_n} \sum_{x \in \mathbf{R}^{-1}(r)} P(x|\phi, l) P(l|\phi) \ . \tag{2}$$

We can approximate the two integrals in (2) by the finite Monte Carlo sums,

$$P(r|\phi) \approx \frac{1}{N} \sum_{j=1}^{N} \frac{1}{M} \sum_{h=1 : x^{(h)} \in \mathbf{R}^{-1}(r)}^{M} P(x^{(h)}|\phi, l^{(j)}),\ l^{(j)} \sim P(l|\phi) \ . \tag{3}$$

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

under ϕ. Therefore, $P(\phi|r) \propto P(r|\phi)P(\phi)$

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

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 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 [7], 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) all nodes are lattice points in $\mathbf{R}^{-1}(r)$ and (2) edges between a node x and a node y are possible $\iff$ $x-y\in\mathcal{M}_{\mathbf{R}}$. 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 [7] implemented in algebraic software packages such as `Macaulay 2` [12] and `4ti2` [13]. 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. Markov bases can also be used to construct implicit graphs of a given radius by expanding from the observed x_o satisfying $\mathbf{R}x_o = r_o$ as described in §4.2.

3 Linear experiments of the SFS

3.1 Number of Segregating Sites

We first describe a classical summary in population genetics termed the *number of segregating sites* and denoted by s [2]. One can express s as a sum of the x_i's:

$$S(x) := \sum_{i=1}^{n-1} x_i = s : \mathcal{X}_n^m \to \mathcal{S}_n^m \ . \tag{4}$$

S is the statistic of the n-coalescent experiment $\mathfrak{X}_{013} := (\mathcal{S}_n^m, \sigma(\mathcal{S}_n^m), \mathcal{P}_{\Phi})$ [1, §3.1 and Fig. 2]. For some fixed sample size n at m homologous and at most biallelic sites, consider the set of SFS that have the same number of segregating

sites s. Then this set denoted by $S^{-1}(s) = \{x \in \mathcal{X}_n^m : S(x) = s\}$ is an s-simplex with cardinality given by the number of compositions of s by $n-1$ parts, i.e.,

$$|S^{-1}(s)| = \binom{s+n-2}{s} = \begin{cases} \frac{\prod_{i=1}^{n-2} s+i}{(n-2)!} & \text{if: } n > 2 \\ 1 & \text{if: } n = 1, s = 0, \text{ or } n = 2 \\ 0 & \text{if: } n = 1, s > 0 \ . \end{cases}$$

The conditional probability of S is Poisson distributed with rate parameter given by the product of the total tree size $l_\bullet := \sum_{i=1}^{n-1} l_i$ and the mutation rate parameter ϕ_1 in ϕ

$$\begin{aligned} P(S = s|\phi, {}^c t) = P(S = s|\phi, l) &= \sum_{x \in S^{-1}(s)} P(x|\phi, l) \\ &= \sum_{x \in S^{-1}(s)} e^{-\phi_1 l_\bullet} (\phi_1 l_\bullet)^s \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i! = e^{-\phi_1 l_\bullet} (\phi_1 l_\bullet)^s / s! \quad (5) \end{aligned}$$

3.2 Heterozygosity

Another classical summary statistic called *average heterozygosity* is also a symmetric linear combination of SFS x [3]. We define *heterozygosity* $Z(x) = z$ and average heterozygosity $\mathit{\Pi}(x) = \pi$ as follows:

$$Z(x) := \sum_{i=1}^{n-1} i(n-i)x_i, \quad \mathit{\Pi}(x) := \frac{1}{\binom{n}{2}} Z(x) \ . \tag{6}$$

Z is the statistic of the n-coalescent experiment $\mathfrak{X}_{012} := (\mathcal{Z}_n^m, \sigma(\mathcal{Z}_n^m), \mathcal{P}_{\Phi})$ [1, §3.1 and Fig. 2]. 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 the integer lattice $\mathcal{X}_n^m$. The conditional probability $P(Z|\phi, a) = P(\mathit{\Pi}|\phi, a) = P(Z = z|\phi, l)$ is

$$P(Z = z|\phi, l) = \sum_{x \in Z^{-1}(z)} P(x|\phi, l) = e^{-\phi_1 l_\bullet} \sum_{x \in Z^{-1}(z)} (\phi_1 l_\bullet)^{\sum_{i=1}^{n-1} x_i} \frac{\prod_{i=1}^{n-1} \bar{l}_i^{x_i}}{\prod_{i=1}^{n-1} x_i!} \ . \tag{7}$$

3.3 Tajima's D

Tajima's D statistic [3] only depends on the number of segregating sites (4), average heterozygosity (6) and the sample size n, as follows:

$$D(x) := \frac{\mathit{\Pi}(x) - S(x)/d_1}{\sqrt{d_3 S(x) + d_4 S(x)(S(x)-1)}} \ , \tag{8}$$

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

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

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 non-negative integer lattice $\mathbb{Z}_+^{n-1}$. Let the intersection of $\{x : \mathbf{R}x = r\}$ with $\mathbb{Z}_+^{n-1}$ be the set:

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

Note that $\mathbf{R}^{-1}(r)$ is the set of site frequency spectra with the same $r = (s, z)'$. Since n is fixed, every SFS x in $\mathbf{R}^{-1}(r)$ has the same s, z, π and d.

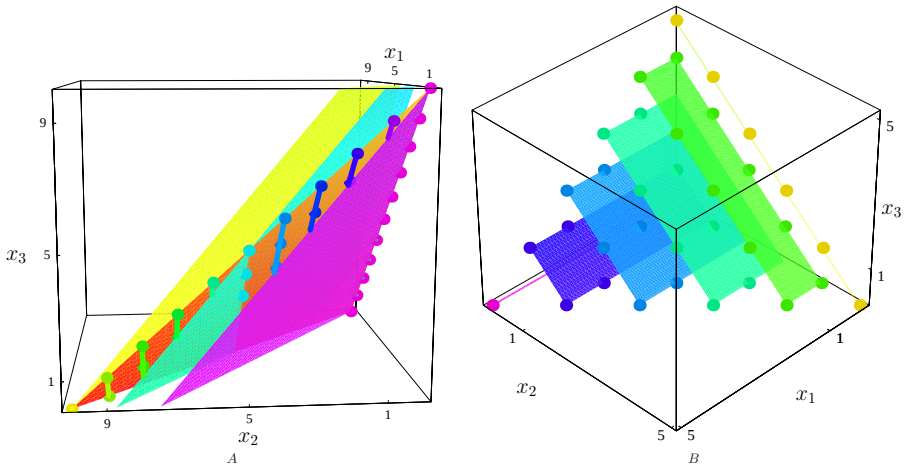

Fig. 1. A: Polytopes containing $\mathbf{R}^{-1}((s, z)')$, where $z \in \{30, 31, \dots, 40\}$, $n = 4$ and $s = 10$ are at the intersection of the s-simplex, $\mathbb{Z}_+^3$ and each of the z-simplexes. B: Projected rectangular polytopes containing $\mathbf{R}^{-1}((s, z)')$, where $z \in \{20, 22, \dots, 30\}$, $n = 5$ 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 the s-simplex given by $x_3 = S - x_1 - x_2$. Fig. 1 A 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 $\mathbf{R}_2 = (3, 4, 3)$, ${\mathbf{R}_2}^{-1}(z) := \{x \in \mathbb{Z}_+^3 : 3x_1 + 4x_2 + 3x_3 = z\}$, the set of SFS satisfying a given z, is the intersection of $\mathbb{Z}_+^3$ with the z-simplex given by $x_3 = (z - 3x_1 - 4x_2)/3$. Fig. 1 A 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 $(x_1, z - 3S, -z + 4S - x_1)$ and satisfy both s and z. In Fig. 1 A, 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. 1 A) 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. 1 A).

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 $0 \leq x_1 \leq 3S - z/2$, $0 \leq x_2 \leq S$. In Fig. 1 B, 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 [14] as implemented in the software package `LattE` [15] to obtain $|\mathbf{R}^{-1}((s, z)')|$ for 1000 data sets simulated under the standard neutral n-coalescent [16] 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 100kbp 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 efficient for smaller m. Our choices of ϕ_1^* and m are biologically motivated by a previous study on human SNP density [17].

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 $y_j := \mathbb{1}_{\{i:i\neq n-i\}}(j)\ x_j + x_{n-j}$, 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. 1 A 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 $r = (10, 36)'$,

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

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

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

and once again we can see from Fig. 1 B 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 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 sate of the nucleotide. It is determined by the map $Y(x) = y : \mathcal{X}_n^m \mapsto \mathcal{Y}_n^m$:

$$\begin{aligned} Y(x) &:= (Y_1(x), Y_2(x), \ldots, Y_{\lfloor n/2 \rfloor}(x)), \\ Y_j(x) &:= x_j \mathbb{1}_{\{i:i\neq n-i\}}(j) + x_{n-j},\ j \in \{1, 2, \ldots, \lfloor n/2 \rfloor\} \end{aligned} \tag{9}$$

Y is the statistic of the n-coalescent experiment $\mathfrak{X}_{011} := (\mathcal{Y}_n^m, \sigma(\mathcal{Y}_n^m), \mathcal{P}_{\Phi})$ [1, §3.1 and Fig. 2]. 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 i-th 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 i-th row $\mathbf{Y}_i$ is:

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

A 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:

$$m_i = e_i - e_{n-i}, \quad i = 1, 2, \ldots, \lfloor n/2 \rfloor \qquad m_{n-i} = -e_i + e_{n-i}, \quad i = 1, 2, \ldots, \lfloor n/2 \rfloor.$$

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 $j \in \{1, 2, \ldots, \lfloor \frac{n-1}{2} \rfloor\}$, (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:

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

3.5 Results

We briefly apply the methods of the previous subsections to estimate the scaled mutation rate ϕ_1^* and the exponential growth rate ϕ_2^* based on the statistics (s, z) and $(s, z, x^{\circledast})$ of the SFS x at one non-recombining locus of length m from n samples. The performance of our estimator is assessed over $1,000$ data sets that were simulated under our standard null model with the locus-specific scaled mutation rate $\phi_1^* = 10.0$, constant population size with the growth rate $\phi_2^* = 0.0$. Our point estimate $(\widehat{\phi_1}, \widehat{\phi_2})$ of (ϕ_1^*, ϕ_2^*) based on (s, z) and $(s, z, x^{\circledast})$ is the maximum *aposteriori* estimate obtained from a histogram estimate of the posterior $P(\phi|x)$ (Table 1). 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 estimators are the result of exactly approximate Bayesian computations (with $\epsilon = 0$) as we integrate exhaustively over all SFS in $\mathbf{R}^{-1}((s, z)')$ when we compute $P(\phi|(s, z))$ or $P(\phi|(s, z, x^{\circledast}))$. However, we have fixed the epoch times at its expectation under ϕ_2, i.e. $\bar{t} = E(t|\phi_2)$, in order to focus on the effect of ignoring the topological information $x^{\circledast}$ when conditioning on (s, z).

	$s, z, \bar{t}$								$s, z, x^{\circledast}, \bar{t}$							
n	$\widehat{\phi_2}$			$\widehat{\phi_1}$			$(\widehat{\phi_1}, \widehat{\phi_2})$		$\widehat{\phi_2}$			$\widehat{\phi_1}$			$(\widehat{\phi_1}, \widehat{\phi_2})$	
n	$\overline{se}$	bs	$c_{99\%}$	$\overline{se}$	bs	$c_{99\%}$	$c_{99\%}$	$Qrt(\hat{\mathcal{K}})$	$\overline{se}$	bs	$c_{99\%}$	$\overline{se}$	bs	$c_{99\%}$	$c_{99\%}$	$Qrt(\hat{\mathcal{K}})$
4	663	18	55.6	797	21	47.3	58.0	$\{0.08, 0.11, 0.21\}$	452	11	61.4	317	10	52.3	61.3	$\{0.08, 0.11, 0.21\}$
6	453	13	55.1	529	14	54.5	52.9	$\{0.10, 0.15, 0.25\}$	276	8	72.1	253	8	58.6	69.2	$\{0.10, 0.15, 0.25\}$
8	339	10	62.0	406	12	55.7	56.4	$\{0.11, 0.18, 0.31\}$	191	5	80.0	201	6	61.8	72.0	$\{0.11, 0.18, 0.31\}$

Table 1. Performance of our estimator of ϕ_1^* and ϕ_2^* based on SFS statistics (s, z) and $(s, z, x^{\circledast})$ (see text).

There are several quantities one can use to gauge the efficiency of our estimator $\widehat{\Phi}$ at this point. Our performance measures based on 1000 replicates of SFS summaries simulated under $\phi_1^* = 10.0$ and $\phi_2^* = 0.0$ can help make natural connections to the theory of approximate sufficiency [11], as we not only measure the bias (bs), root-mean-squared-error ($\sqrt{\overline{se}}$) 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. Notice the increase in the joint empirical coverage as the statistic used to estimate ϕ^* changes from (s, z) to $(s, z, x^{\circledast})$ (also compare to [1, Table 1] that uses the entire SFS x). This estimator can be improved by importance sampling epoch times at the expense of great computational cost (see §5).

Due to the controlled Markov chain $\{F^{\downarrow x^{\circledast}}(k)\}_{k\in[n]_+}$ [1, Prop. 9], it is straightforward to directly obtain the posterior from the much finer resolution of the observed SFS x, as opposed to integrating over a large set of SFS that exactly satisfy few of its linear summaries. Nonetheless, these computations over population genetic fibers shed algebraic insights and provide rigorous and exact benchmarks to compare, correct and improve simulation-intensive ABC/ALC algorithms in the current molecular population genetic literature that ignore topological information in the hidden space of genealogies. Finally, it is important to bear in mind, especially when comparing to simulation-intensive ABC/ALC methods in the literature that we are purposely using statistics 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 intra-locus 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 non-recombining loci.

4 Natural Extensions of Markov Basis Methods

4.1 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 $y_1 := x_1 + x_{(n-1)}$ [4] and Fay and Wu's $\theta_H := (n(n-1))^{-1} \sum_{i=1}^{n-1} (2\ i^2\ x_i)$ [5].

4.2 Neighborhoods of Site Frequency Spectra

Finally, one need not restrict the summaries to linear combinations of the SFS. Such methods are naturally amenable to linear combinations of finer summaries of the full data that are more general than the SFS.

In summary, 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 the $|\mathcal{M}_{\mathbf{R}}|$. 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 and thus render convergence diagnostics extremely heuristic. 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:

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

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 this can facilitate in the exact inference from linear summaries of the SFS.

4.3 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. In fact, linear combinations of any summary of the full data d_o may be used in such an exactly ABC scheme. For example, consider data from two known sub-populations 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 sub-populations. Unlike the full SFS $x \in \mathbb{Z}_+^{n-1}$,

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

where, x_i^J is the number of sites that have i samples only from sub-population $J \in \{A, B\}$ sharing a mutation (there are no mutations at these sites in the other sub-population). We can think of x^A and x^B as sub-population 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:

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

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.,

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

Note that the total number of segregating sites

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

We are interested in the sub-population determined SFS $\ddot{x}$ given by,

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

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.

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

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

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

Observe that Tajima's D for the entire sample as well as the sub-population specific D^A and D^B computed from the sites that are segregating only within sub-population 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 $\mathbf{R}(r)$ 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" ($\overline{l} \in \triangle_{n-2}$). 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 sub-population A alone ($l_\bullet^A$), in sub-population B alone ($l_\bullet^B$) and those leading to mutations in both sub-populations ($l_\bullet^{AB}$). 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 sub-population-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 sub-population A. Next we compress these labeled lineages into the relative time leading to singletons, doubletons, ..., "n^A-tons" exclusively within sub-population A. These labeled relative times yield $\overline{l}^A \in \triangle_{n^A-1}$. By an analogous labeling and compression of ${}^c\ddot{t}$ we obtain $\overline{l}^B \in \triangle_{n^B-1}$. Finally, we obtain the probabilities $\overline{l}^{AB} \in \triangle_{n-3}$ by labeling the lineages on ${}^c\ddot{t}$ that lead to both sub-populations.

5 Discussion

In this article we discussed applications of Markov bases to population genetic statistics via the unlabeled n-coalescent and the associated controlled Markov

chain $\{F^{\lfloor x^{\circledast}}(k)\}_{k\in[n]_+}$ developed in [1]. We used computer software to compute each Markov basis. Unfortunately, the number of elements in a Markov basis gets exponentially large if we increase the sample size n. As discussed in [18], most of the elements in a Markov basis are used for sparse vector r. Thus, for some cases if we assume that the right-hand-side r is positive then a subset of a Markov basis whose elements connect all points in a fiber has much smaller elements than those in a Markov basis. Thus it is interesting and practical to find a subset of a Markov basis which connect all points in $\mathbf{R}^{-1}(r)$ if we assume $r > 0$.

The interesting aspect of our population genetic fibers is the combinatorial complexity of the likelihood functions over them. Each of these *combinatorially many* f-sequence-indexed likelihood functions that we need to integrate over our fiber are uniquely and highly structured leading to f-sequence-specific "islands" and "peninsulars" of fiber subsets surrounded by zeros of the likelihood. This is in contrast to the *single* exponentially concentrated likelihood function in fibers that typically arise in contingency table problems. Nonetheless, the irreducibility of the fiber guaranteed by a Markov basis helps when integrating over the f-sequence-indexed likelihood functions, especially using ideas in §4.2. A naive application of MCMC methods, including tempered methods, Population MCMC and static particle filters of the sequential importance sampling family, to obtain a Monte Carlo estimate of the integral over our fibers, leads to highly biased estimates due to the distinctness of the MCMC convergence bottle-necks imposed by each of the f-sequence-indexed likelihood functions.

The exact ABC/ALC algorithms in this article, that are based on a few coarse linear summaries of the SFS and ensure a zero acceptance radius ($\epsilon = 0$), are computationally less efficient and provide less information when compared to the methods in [1], that are not only based on the entire SFS but also ensure $\epsilon = 0$. 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.

ACKNOWLEDGMENTS

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 Peter Donnelly during the course of this study. R.S. thanks Celine Becquet for discussions on summary statistics of structured populations. R.Y. was supported by NIGMS grant 1R01GM086888-01.

References

1. Sainudiin, R., Thornton, K., Griffiths, R., McVean, G., Donnelly, P.: Coalescent experiments I: Unlabeled n-coalescent and the site frequency spectrum – UCDMS Research Report 2009/7, may 19, 2009 (submitted). Available at `http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/CoalExpsI.pdf` (2009)

2. Watterson, G.: On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7** (1975) 256–276
3. Tajima, F.: Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics **123** (1989) 585–595
4. Hudson, R.: The how and why of generating gene genealogies. In Clark, A., Takahata, N., eds.: Mechnisms of Molecular Evolution. Sinauer (1993) 23–36
5. Fay, J., Wu, C.: Hitchhiking under positive darwinian selection. Genetics **155** (2000) 1405–1413
6. Wakeley, J.: Coalescent Theory: An Introduction. Roberts & Co. (2007)
7. Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. Annals of Statistics **26** (1998) 363–397
8. Weiss, G., von Haeseler, A.: Inference of population history using a likelihood approach. Genetics **149** (1998) 1539–1546
9. Beaumont, M., Zhang, W., Balding, D.: Approximate Bayesian computation in population genetics. Genetics **162** (2002) 2025–2035
10. Marjoram, P., Molitor, J., Plagnol, V., Tavare, S.: Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **100** (2003) 15324–15328
11. Cam, L.L.: Sufficiency and approximate sufficiency. Ann. Math. Stats. **35** (1964) 1419–1455
12. Grayson, D., Stillman, M.: `Macaulay 2`, a software system for research in algebraic geometry. Available at `www.math.uiuc.edu/Macaulay2` (2004)
13. Hemmecke, R., Hemmecke, R., Malkin, P.: 4ti2 version 1.2—computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. Available at www.4ti2.de (2005)
14. Barvinok, A.: Polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. Math. of Operations Research **19** (1994) 769–779
15. Loera, J.D., Haws, D., Hemmecke, R., Huggins, P., Tauzer, J., Yoshida, R.: `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` (2004)
16. Hudson, R.: Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18** (2002) 337–338
17. Sainudiin, R., Clark, A., Durrett, R.: Simple models of genomic variation in human SNP density. BMC Genomics **8** (2007) 146
18. Chen, Y., Dinwoodie, I., Yoshida, R.: Markov chains, quotient ideals, and connectivity with positive margins. Algebraic and Geometric Methods in Statistics dedicated to Professor Giovanni Pistone (P. Gibilisco, E. Riccomagno, M.-P. Rogantin, H.-P. Wynn, eds.) (2008)