Efficiently inferring the demographic history of many populations with allele count data

John A. Kamm[1,2], Jonathan Terhorst[3], Richard Durbin[1,2], and Yun S. Song[*4,5,6]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
[2]Department of Genetics, University of Cambridge, Cambridge, UK
[3]Department of Statistics, University of Michigan, Ann Arbor, USA
[4]Computer Science Division, University of California, Berkeley, USA
[5]Department of Statistics, University of California, Berkeley, USA
[6]Chan Zuckerberg Biohub, San Francisco, USA

March 22, 2018

Abstract

The sample frequency spectrum (SFS), or histogram of allele counts, is an important summary statistic in evolutionary biology, and is often used to infer the history of population size changes, migrations, and other demographic events affecting a set of populations. The expected multipopulation SFS under a given demographic model can be efficiently computed when the populations in the model are related by a tree, scaling to hundreds of populations. Admixture, back-migration, and introgression are common natural processes that violate the assumption of a tree-like population history, however, and until now the expected SFS could be computed for only a handful of populations when the demographic history is not a tree. In this article, we present a new method for efficiently computing the expected SFS and linear functionals of it, for demographies described by general directed acyclic graphs. This method can scale to more populations than previously possible for complex demographic histories including admixture. We apply our method to an 8-population SFS to estimate the timing and strength of a proposed "basal Eurasian" admixture event in human history. We implement and release our method in a new open-source software package `momi2`.

1 Introduction

All natural populations undergo evolutionary processes of migration, size changes, and divergence, and the history of these demographic events shape their present genetic diversity. Thus, inferring demographic history is of central concern in evolutionary and population genetics, both for its intrinsic interest (e.g., in dating the out-of-Africa migration of modern humans (Schaffner et al., 2005; Gutenkunst et al., 2009)) and also for biological applications (such as distinguishing the effects of natural selection from demography (Beaumont and Nichols, 1996; Boyko et al., 2008)). However, genetic sequence data and the space of possible demographic models are both very high dimensional objects, leading to numerous statistical and computational challenges when inferring demographic history from genetic data.

The joint sample frequency spectrum (SFS) is the multidimensional histogram of mutant allele counts in a sample of DNA sequences, and is a popular summary statistic which lies at the core of thousands of

*To whom correspondence should be addressed: yss@berkeley.edu

empirical studies in population genetics; e.g., see Wakeley and Hey (1997); Griffiths and Tavaré (1998); Nielsen (2000); Gutenkunst et al. (2009); Coventry et al. (2010); Gazave et al. (2014); Gravel et al. (2011); Nelson et al. (2012); Excoffier et al. (2013); Jenkins et al. (2014); Bhaskar et al. (2015); Jouganous et al. (2017). Recently, progress has also been made on theoretical fronts to characterize statistical properties of SFS-based inference. In particular, studies of identifiability (Myers et al., 2008; Bhaskar and Song, 2014) and the rate of convergence (Terhorst and Song, 2015; Baharian and Gravel, 2018) have been carried out.

Demographic history can be inferred by fitting the observed value of the SFS to its expected value in a composite likelihood framework. The expected SFS can be efficiently computed when the demographic history is a tree, and in previous work we developed a method *momi* to compute the SFS of hundreds of populations related by a tree (Kamm et al., 2017). However, natural populations are often related by a more complex history that is not tree-like, as gene flow (the exchange of migrants between populations) adds extra edges to the topology associated with the demographic history. In this case, computing the expected SFS is much more computationally demanding, and existing methods for computing the exact expected SFS can scale to only a handful of populations (Gutenkunst et al., 2009; Jouganous et al., 2017).

In this article, we extend our previous algorithm *momi* to handle discrete (or pulse) migration events between populations, in which case demographies are described by general directed acyclic graphs (DAGs). Our new method `momi2` can compute the *exact* expected SFS with admixture for more populations than previously possible, and uses novel insights from a stochastic process known as the Lookdown Construction (Donnelly and Kurtz, 1996; Donnelly et al., 1999). In addition, `momi2` utilizes automatic differentiation (Corliss et al., 2002; Bhaskar et al., 2015) to compute gradients of the SFS, which we use to efficiently search the parameter space during optimization. Finally, `momi2` can efficiently compute linear functionals of the SFS, which we exploit to compute the expected values of a number of standard population genetic summary statistics under complex demography.

The rest of this paper as organized as follows. In Section 2 we provide some background and survey related work. Section 3 describes our method, first with an illustrative example in Section 3.1, and then with formulas and pseudocode in Section 3.2. Finally, in Section 4, we apply our method to an 8-population SFS, including ancient and contemporary human populations, to estimate the timing and strength of a proposed "basal Eurasian" admixture event in human history. We defer all proofs to the Appendix.

2 Background

Suppose a sample of $\mathbf{n} = (n_1, \dots, n_D)$ genomes have been sampled from D "demes" or populations. The positions in the genome where the samples are not all identical are called *segregating sites*. In most organisms mutations are rare; most sites are not segregating. It is therefore reasonable to assume, as we do from now on, that each position in the genome has experienced at most a single mutation in its history, and that each individual can be labeled as having the "ancestral" or "derived" (mutant) allele at each segregating site. In population genetics, this simplifying assumption is known as the *infinite sites model*.

The *sample frequency spectrum* (SFS) is a D-dimensional array $[\mathrm{f}_{\mathbf{x}}] \in \mathbb{Z}^{(n_1+1)\times\cdots\times(n_D+1)}$ whose entry $\mathrm{f}_{\mathbf{x}}$ counts the number of segregating sites with exactly $\mathbf{x}$ copies of the derived allele and $\mathbf{n} - \mathbf{x}$ copies of the ancestral allele, where $\mathbf{x} = (x_1, \dots, x_D) \in \mathbb{N}_0^D$ with $0 \le x_d \le n_d$ for each $d = 1, \dots, D$. Note we only consider segregating sites with 2 alleles, so $\mathrm{f}_{\mathbf{0}} = \mathrm{f}_{\mathbf{n}} = 0$ by definition. Compared to the full data set (i.e., the complete genetic sequences of all $n = n_1 + \cdots + n_d$ genomes), the SFS $[\mathrm{f}_{\mathbf{x}}]$ is a compressed, low-dimensional summary which nevertheless preserves much of the signal about the various population size changes, divergence times, and admixture events that occurred over the course of the populations' history.

2.1 Demographic events

The expected multipopulation SFS can be obtained by integrating over random genealogies formed by a backwards-in-time stochastic process known as the *structured coalescent* (Kingman, 1982; Takahata, 1988; Notohara, 1990). Before getting to the technical details, we first review the basic dynamics of this process in order to build intuition for how the data have power to infer population splits, size changes, and gene flow events. See Durrett (2008) for a more detailed introduction.

Informally, the topology and branch lengths of genealogies are affected by a demographic history in two ways:

1. Two lineages may not coalesce into a common ancestor until they reside in the same population, and the time until this occurs is affected by migration patterns and population split times.

2. At any particular point in time, two members within the same population are more likely to have a common parent if the population size is small; so, for example, residents of a small village will typically be more closely related than residents of a large city.

Regarding the second point, we define the *scaled effective population size* $\eta(t)$ such that the rate at which any two lineages find a common ancestor at time t is $1/\eta(t)$. Under the simplest random mating model (the Wright Fisher model, cf. Durrett (2008)), the census population size exactly equals $T\eta$, where T is the number of generations per unit time; more generally, η scales with the number of breeding individuals in the population. Thus, estimating η allows us to infer size change events such as bottlenecks, exponential growth, and population crashes.

If we could observe the true distribution of genealogies, then we could directly infer demographic history from the waiting times between coalescence events, following the principles listed in items 1 and 2 above. However, since genealogies are never directly observed, we must make inferences about demographic history indirectly using mutation data.

2.2 Likelihoods and the site frequency spectrum

Consider a genome with L positions and mutation rate $\frac{\theta}{L}$ per position. So at any given position in the genome, mutations arise on the tree there as a Poisson point process with rate $\frac{\theta}{L}$. The chance of 2 or more mutations at a single position is $O(\frac{1}{L^2})$, and taking the limit $L \to \infty$ we arrive at the aforementioned *infinite sites approximation* (Kimura, 1969; Durrett, 2008), which assumes that each segregating site was caused by a single mutation, so that each allele may be labeled as ancestral or derived.

The observed segregating sites are not independent, because trees at neighboring positions are correlated. Unfortunately, even in the simplest case of a single population with constant size, an analytic expression for the likelihood of mutation data at a set of linked (non-independent) sites is not known (Bhaskar et al., 2015). Therefore, SFS data are generally used with composite likelihood methods. Recall that $\mathrm{f}_{\mathbf{x}}$ is the total number of segregating sites with derived allele count pattern $\mathbf{x} = (x_1, \ldots, x_D)$. Define $\phi_{\mathbf{x}} = \frac{1}{\theta}\mathbb{E}[\mathrm{f}_{\mathbf{x}}]$, i.e., $\phi_{\mathbf{x}}$ is the expected frequency of $\mathbf{x}$ per unit mutation rate. Equivalently, $\phi_{\mathbf{x}}$ is the expected branch length subtending $\mathbf{x}$ leaves in a random coalescent tree (i.e., with x_1 descendants in population 1, x_2 in population 2, and so on). A commonly used composite likelihood is the *Poisson random field* model (Sawyer and Hartl, 1992), which assumes that the total number of segregating sites is Poisson with rate $\theta \sum_{\mathbf{x}} \phi_{\mathbf{x}}$, and that the patterns at the observed sites are independent with sampling probabilities proportional to $\phi_{\mathbf{x}}$; this yields a log-likelihood of

$$\hat{\mathcal{L}} \propto \|\mathrm{f}\|_1 \log(\theta\|\phi\|_1) - \theta\|\phi\|_1 + \sum_{\mathbf{x}} \mathrm{f}_{\mathbf{x}} \log \frac{\phi_{\mathbf{x}}}{\|\phi\|_1}, \tag{1}$$

where $\|\mathrm{f}\|_1 := \sum_{\mathbf{x}} \mathrm{f}_{\mathbf{x}}$ and $\|\phi\|_1 := \sum_{\mathbf{x}} \phi_{\mathbf{x}}$. Demographic history can then be inferred by searching for the parameter values that maximize $\hat{\mathcal{L}}$.

2.3 Existing work and our contribution

The composite log-likelihood $\hat{\mathcal{L}}$ in (1) requires us to compute $\phi_{\mathbf{x}}$, the expected branch length subtending $\mathbf{x}$. Let G be a random genealogical tree sampled under the demography, and $L_{\mathbf{x}}(G)$ be the total length of all branches in G subtending $\mathbf{x}$, so that

$$\phi_{\mathbf{x}} = \mathbb{E}[L_{\mathbf{x}}(G)] = \int_G L_{\mathbf{x}}(G)\, d\mathbb{P}(G). \tag{2}$$

The integral (2) is difficult since the support of G is at least as large as the number of labeled binary trees with n leaves, a quantity which grows faster than exponentially in the sample size n. Consequently, a number of different methods have been proposed for evaluating (or approximating) the integral (2). Sampling-based methods include Markov chain Monte Carlo (Griffiths and Tavaré, 1997; Nielsen, 2000), importance sampling (Stephens and Donnelly, 2000; De Iorio and Griffiths, 2004), simulation (Excoffier and Foll, 2011; Excoffier et al., 2013), and ABC (Wegmann et al., 2010). The main advantage of these methods is their flexibility: since, as mentioned above, computing the likelihood of the data is trivial conditional on G, these methods can be used on a rich class of models. However, the high dimension of $\phi_{\mathbf{x}}$ makes it impractical to compute by sampling unless $\mathcal{D}$ is small. In particular, since the support of $\mathbf{x}$ grows like $O(n^{\mathcal{D}})$, Monte Carlo methods will assign zero mass to configurations that are actually observed in the data.

A second approach, implemented in the software *∂a∂i* (Gutenkunst et al., 2009), computes $\phi_{\mathbf{x}}$ by numerically solving PDEs arising from the Wright-Fisher diffusion (Ewens, 2004), which is dual to the coalescent process described above. For $\mathcal{D}$ populations, this involves numerically solving a $\mathcal{D}$-dimensional integral. The initial *∂a∂i* method in Gutenkunst et al. (2009) could handle up to $\mathcal{D} = 3$ populations; subsequent improvements (Lukić and Hey, 2012; Jouganous et al., 2017) extended this to $\mathcal{D} = 4$ and then $\mathcal{D} = 5$ populations by using spectral representations or alternative basis functions for solving the PDEs.

The third approach for computing $\phi_{\mathbf{x}}$, which includes our method, integrates over the sample allele frequencies "backwards-in-time", exploiting conditional independence relationships to reduce computation. This involves considering the alleles of the sample's ancestors at different points in the demographic history, and integrating out these random variables via inference algorithms for probabilistic graphical models (Pearl, 1982; Felsenstein, 1981; Lauritzen and Spiegelhalter, 1988; Koller and Friedman, 2009). Bryant et al. (2012) and Chen (2012) computed the SFS for finite- and infinite-sites models using this backward-in-time approach under the coalescent. De Maio et al. (2015) and Kamm et al. (2017) substantially lowered the computational burden of this approach by replacing the coalescent with the *continuous-time Moran model* (Durrett, 2008), a stochastic process which induces the same sampling distribution as the coalescent, but using a much smaller state space. However, until now these Moran-based approaches have been limited to analyzing tree-shaped demographies without admixture between populations.

The main contribution of this paper is to extend our previous Moran-based method (Kamm et al., 2017) to allow for demographies defined on general directed acyclic graphs (DAGs), thus allowing for admixture between populations. We describe and implement an algorithm for computing the expected infinite-sites SFS under the multipopulation Moran model with size changes, exponential growth, population splits, and point admixture events (i.e., instantaneous migration "pulses"). This substantially enlarges the space of demographies for which the expected SFS can be accurately computed.

Additionally, our algorithm computes not only individual SFS entries, but also linear functionals of the expected SFS. Specifically, our method computes rank-1 tensor products of the SFS in the same time as a single entry (general linear functionals are sums of these rank-1 products). A number of widely-used statistics in population genetics can be expressed as SFS functionals, and to our knowledge our method is the first to compute expectations of these statistics under complex demography.

We demonstrate our method by using it to infer the history of eight human subpopulations that have undergone multiple admixture events. We complement our theoretical contributions with an open-source,

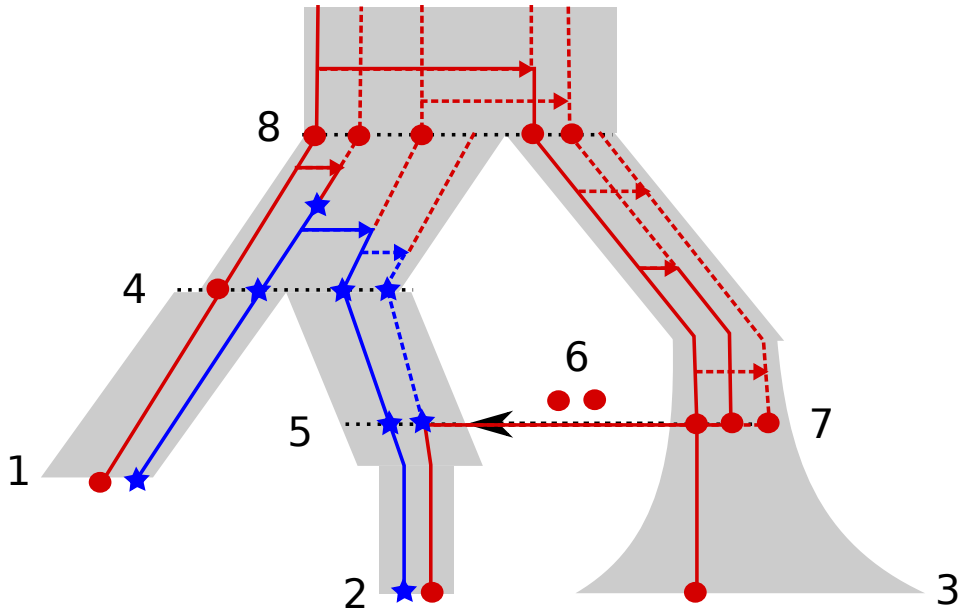
Symbol	Description	Reference
$\mathcal{D}$	Sampled populations are $\{1, \dots, \mathcal{D}\}$	§2
$\mathbf{n}$	# of samples per population $(n_1, \dots, n_{\mathcal{D}})$	"
$\mathrm{f}_{\mathbf{x}}$	# of sites with $\mathbf{x}$ derived and $\mathbf{n} - \mathbf{x}$ ancestral alleles	"
θ	Mutation rate	§2.2
$\phi_{\mathbf{x}}$	Expected branch length with $\mathbf{x}$ descendants, $\frac{1}{\theta}\mathbb{E}[\mathrm{f}_{\mathbf{x}}]$	"
$\mathcal{G}$	Population graph	Figure 1b
v	A vertex in $\mathcal{G}$, corresponding to a population	§3.2
τ_v	time between top and bottom events of v	"
$\eta_v(t)$	scaled population size of v at $t \in [0, \tau_v]$	"
n_v	# of samples with ancestry in v	"
$X_v^{(t)}$	# of derived alleles in v at time t	"
$\mathbf{X}_S^{(\mathbf{t})}$	Vector of allele counts $(X_{v_1}^{(t_1)}, \dots, X_{v_k}^{(t_k)})$ at $S = \{v_1, \dots, v_k\}$	"
$X_v, \mathbf{X}_S$	Shorthand for $X_v^{(0)}, \mathbf{X}_S^{(\mathbf{0})}$ respectively	"
$\mathcal{T}$	Event tree	Figure 1c
E	A join, split or leaf event in $\mathcal{T}$	"
K_E	Populations we are keeping track of at E	"
$\ell_{\mathbf{x},\mathbf{z}}^{E,\mathbf{t}}$	Conditional likelihood $\mathbb{P}(\mathbf{X}_{\mathtt{Leaves}(E)} = \mathbf{z} \mid \mathbf{X}_{K_E}^{(\mathbf{t})} = \mathbf{x})$	Eq. (4)
$\phi_{\mathbf{z}}^E$	$\mathbb{E}$[branch length at/below K_E with $\mathbf{z}$ descendants]	Eq. (5)
$\ell_{\mathbf{x}}^{E,\mathbf{t}}, \phi^E$	Shorthand for $\ell_{\mathbf{x},\mathbf{z}}^{E,\mathbf{t}}, \phi_{\mathbf{z}}^E$ respectively when $\mathbf{z}$ fixed	"
n_{tot}	The total number of sampled lineages $\sum_{d=1}^{\mathcal{D}} n_d$	Appendix A.1
$\mathcal{M}_{v,t,(i)}$	Label and allele of the ith lineage in v at $t \in [0, \tau_v]$	"
$\mathcal{M}_{v,t}$	$(\mathcal{M}_{v,t,(1)}, \mathcal{M}_{v,t,(2)}, \dots)$	"
$\mathcal{M}_{E,\mathbf{t}}$	$(\mathcal{M}_{v_1,t_1}, \dots, \mathcal{M}_{v_k,t_k})$ where $K_E = \{v_1, \dots, v_k\}$ and $\mathbf{t} = (t_1, \dots, t_k)$	"
$X_{v,m}^{(t)}$	the # derived among the 1st m lineages in v, t. (Note: $X_v^{(t)} \equiv X_{v,n_v}^{(t)}$)	"

Table 1: Notation. Rows in the second part of the table are only used for proofs in the Appendix and may be ignored in the main text.

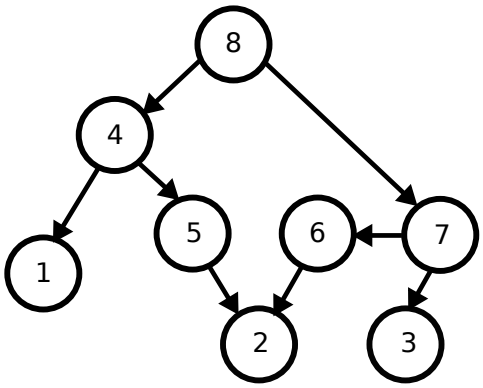
user-friendly software implementation that will enable practitioners to deploy our method. The software uses automatic differentiation (Corliss et al., 2002; Bhaskar et al., 2015; Maclaurin et al., 2015) to compute derivatives of the SFS, leading to efficient optimization and parameter inference. Our package, called `momi2`, is available for download at `https://github.com/popgenmethods/momi2`.

3 Method

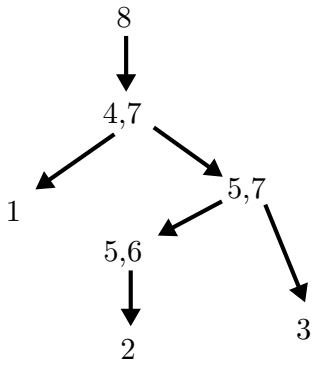
In this section, we describe the algorithm implemented in `momi2` for computing the expected SFS under complex demographies. We begin in Subsection 3.1 with an illustrative example that highlights the novel aspects of our work. Then in Subsection 3.2 we provide pseudocode for our algorithm, and state the formulas used by our algorithm as Propositions. The proofs of these Propositions, and the proof for the correctness of our algorithm, requires substantial additional notation, and we defer this to Appendix A.2. For ease of reference, the symbols we use are summarized in Table 1.



(a) An example of a 3-population Moran model. The bottom of the graph corresponds to the present and the top to the past. Population 2 receives admixture from population 3 after splitting from population 1. Other features of the demography include archaic samples in population 1, and various size changes along the edges of this demography.



(b) The corresponding population DAG $\mathcal{G}$. Each vertex corresponds to a collection of alleles in (a).



(c) The corresponding event tree $\mathcal{T}$ is a junction tree of the DAG in $\mathcal{G}$ (b); each vertex of $\mathcal{T}$ corresponds to a set of vertices from $\mathcal{G}$.

Figure 1: An illustrative example of a 3-population Moran model (a), along with (b) the corresponding DAG $\mathcal{G}$, and (c) event tree $\mathcal{T}$.

3.1 Example

Consider the model depicted in Figure 1. This model has 3 sampled (leaf) populations, related as follows. Populations 1 and 2 are sisters, with Population 3 an outgroup to them; however a pulse of migrants from 3 to 2 occurs after the split between populations 1 and 2.

At the leaves, we observe $(n_1, n_2, n_3) = (2, 2, 1)$ samples, with $(X_1, X_2, X_3) = (1, 1, 0)$ copies of the derived (blue star) allele. We wish to compute the expected number of mutations with pattern (X_1, X_2, X_3); to do this, we will integrate over the unobserved variables $(X_4, X_5, X_6, X_7, X_8)$ which represent allele counts at certain internal positions within the demography. The random variables $X_1, \ldots, X_8$ are related to each other by the DAG $\mathcal{G}$ in Figure 1b, with an edge from population v to w if alleles may pass directly from v

into w.

In the realization of Figure 1, the hidden blue allele counts are $(X_4, X_5, X_6, X_7, X_8) = (3, 2, 0, 0, 0)$. The blue mutation occurs in the edge above X_4, spreading to 3 of the lineages. One copy of the blue allele moves on to X_1, while 2 copies move on to X_5. However, due to the admixture, X_2 only inherits 1 blue allele from X_5, inheriting a red allele from the 2 red alleles at X_6.

Under the infinite-sites assumption described above, the mutation observed at this site arose at a single point in the genealogical tree depicted in Figure 1. To compute the expected number of mutations with $(X_1, X_2, X_3) = (1, 1, 0)$, we may condition on the population (i.e., the edge in Figure 1) on which it arose:

$$\begin{aligned}\phi_{1,1,0} &= \frac{1}{\theta}\mathbb{E}[\#\text{ mutations with } (X_1, X_2, X_3) = (1, 1, 0)] \\ &= \frac{1}{\theta}\sum_{v=1}^{8}\sum_{i=1}^{n_v}\mathbb{E}[\#\text{ mutations in population } v \text{ yielding } X_v = i] \\ &\qquad \times \mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0 \mid \text{mutation at } v \text{ with } X_v = i). \end{aligned} \tag{3}$$

We call $\frac{1}{\theta}\mathbb{E}[\#\text{ mutations at } v \text{ with } X_v = i]$, the "truncated SFS"; this only concerns events within a single population and can be computed using the method `momi1`. We refer the interested reader to (Kamm et al., 2017) for further details.

The second term $\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0 \mid \text{mutation at } v \text{ with } X_v = i)$ gives the conditional likelihood of observing the data given a mutation and its allele count at population v. In `momi1`, we computed this term in the case where $\mathcal{G}$ is a tree without admixture using the sum-product (also known as belief propagation) algorithm (Felsenstein, 1981; Koller and Friedman, 2009).

The main result of this work is to extend our previous dynamic program to the case where $\mathcal{G}$ is given by a DAG due to admixture. We will use a dynamic program that is essentially a kind of junction tree algorithm (Koller and Friedman, 2009). This algorithm works by decomposing a DAG graph into a tree in such a way that vertices in the tree correspond to collections of nodes in the original graph. Belief propagation is then applied to the tree decomposition.

We illustrate our algorithm using the example demography in Figure 1c. We call the tree decomposition $\mathcal{T}$ an *event tree*, because each internal node corresponds to either an admixture or split event.[1] We construct the event tree $\mathcal{T}$, and compute the conditional likelihoods $\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0 \mid \text{mutation at } v \text{ with } X_v = i)$, as follows:

1. We initially start with a collection of 3 singleton sets of leaf populations: $\{\{1\}, \{2\}, \{3\}\}$. For $v = 1, 2, 3$, we also keep track of conditional likelihoods of the data beneath v; since we are at the leaves, these are simply the Kronecker delta functions

$$\mathbb{P}(X_v = i \mid X_v = j) = \delta_{i,j} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

2. Going back in time, the first event is the admixture of populations 5 and 6 into 2. To process this event, we remove the set $\{2\}$ and replace it with $\{5, 6\}$; the current collection of sets becomes $\{\{1\}, \{3\}, \{5, 6\}\}$. We make the new set $\{5, 6\}$ the parent of the removed set $\{2\}$ in $\mathcal{T}$.

   In addition, we compute $[\mathbb{P}(X_2 \mid X_5 = i, X_6 = j)]_{i,j}$, the conditional likelihood of the data beneath $\{5, 6\}$ given the allele counts X_5, X_6. We obtain this by applying Lemmas 1 and 2 to the previous conditional likelihood $[\mathbb{P}(X_2 \mid X_2 = i)]_i$. Specifically, we apply Lemma 1 to "lift" the

[1] We omit certain graph preprocessing steps (moralization and triangulation) from the general junction tree algorithm which are not necessary in our setting.

conditional likelihood at population 2 to the time point immediately below the admixture event; then, we apply Lemma 2 to obtain $\mathbb{P}(X_2 \mid X_5, X_6)$, the conditional likelihood at $\{5,6\}$ immediately above the admixture event.

3. The next event has the alleles from 6 splitting off from population 3. To process this event, we merge the clusters containing the relevant populations ($\{5,6\}$ and $\{3\}$); then we remove 3,6 and replace them with their parent population 7, to obtain $\{5,7\}$ as the parent cluster of $\{5,6\}$ and $\{3\}$ in $\mathcal{T}$. After this stage, our collection of sets becomes $\{\{1\},\{5,7\}\}$.

   To obtain $\mathbb{P}(X_2, X_3 \mid X_5, X_7)$, the conditional likelihood of the data at the leaves beneath $\{5,7\}$, we apply Lemma 3 to $\mathbb{P}(X_3 \mid X_3)$ and $\mathbb{P}(X_2 \mid X_5, X_6)$. Lemma 3 computes the conditional likelihood at a split event when the children fall into separate clusters beneath the split (in this case, the child clusters are $\{3\}$ and $\{5,6\}$).

4. The next event is similar, with populations 1 and 5 merging into population 4. After combining the clusters $\{1\}$ and $\{5,7\}$, removing the merged populations 1 and 5, and adding in their parent 4, we are left with the collection $\{\{4,7\}\}$.

   Similar to previous steps, we compute $\mathbb{P}(X_1, X_2, X_3 \mid X_4, X_7)$, the conditional likelihood of the data at the leaves beneath $\{4,7\}$, by applying Lemma 1 to "lift" the conditional likelihoods $\mathbb{P}(X_1 \mid X_1)$ and $\mathbb{P}(X_2, X_3 \mid X_5, X_7)$ to the point immediately below the split event, and then apply Lemma 3 to compute the conditional likelihood at a split event above two independent clusters ($\{1\},\{5,7\}$).

5. The final event has populations 4 and 7 merging into population 8. We replace the cluster $\{4,7\}$ with its parent cluster $\{8\}$ at the root of $\mathcal{T}$.

   To compute $\mathbb{P}(X_1, X_2, X_3 \mid X_8)$ from the child likelihoods $\mathbb{P}(X_1, X_2, X_3 \mid X_4, X_7)$, we first apply Lemma 1 to lift the likelihoods immediately below the split event, and then apply Lemma 4, which computes the conditional likelihood at the parent of a split when the children belong to the same cluster ($\{4,7\}$ in this case).

Finally, at each cluster $\{v_1, v_2, \ldots\}$ in $\mathcal{T}$ with leaves $\{l_1, l_2, \ldots\}$, we have computed $\mathbb{P}(X_{l_1}, X_{l_2}, \ldots \mid X_{v_1}, X_{v_2}, \ldots)$, the conditional likelihoods at the leaves beneath $\{v_1, v_2, \ldots\}$. But to apply equation (3), we need $\mathbb{P}((X_1, X_2, X_3) = (1,1,0) \mid \text{mutation at } v \text{ with } X_v = i)$. This is given by

$$\begin{aligned}
&\mathbb{P}((X_1, X_2, X_3) = (1,1,0) \mid \text{mutation at } v \text{ with } X_v = i) \\
&\quad = \begin{cases} \mathbb{P}((X_1, X_2, X_3) = (1,1,0) \mid X_4 = i, X_7 = 0), & \text{if } v = 4, \\ \mathbb{P}((X_1, X_2, X_3) = (1,1,0) \mid X_8 = i), & \text{if } v = 8, \\ 0, & \text{else,} \end{cases}
\end{aligned}$$

since we assume that each site experiences at most a single mutation in its history, and the only way derived alleles are observed in leaf populations 1,2 is if the corresponding mutation occurs in population 4 or 8.

Remark. The observant reader may have noticed that in Figure 1, population 8 has only $n_8 = 5$ alleles, despite its children having $n_4 + n_7 = 7 > 5$ alleles in total. This is due to the fact that there are only 5 alleles at the leaves, so there are at most 5 ancestors in any population at any point; thus, when populations 4 and 7 merge into population 8, we can drop two of the tracked lineages. To show this formally, we use a stochastic process called the *lookdown construction*, which is a version of the Moran model with a countably infinite number of lineages. However, this analysis requires a great deal of additional notation and is not essential to the remainder of the text, so we defer it to Appendix A.1.

3.2 Algorithms and formulas

We now describe the algorithm to construct the event tree $\mathcal{T}$. We assume that the population graph $\mathcal{G}$ has two types of topological events:

1. Population split: two child populations u, v split from each other; their parent population is w. Looking backward in time, u, v merge and become the population w.

2. Population admixture: a single child population u, inherits from exactly two parent populations v, w, with the probability that an allele comes from v (w, respectively) being p ($1-p$, respectively).

Note that more complicated events, such as trifurcating splits or symmetric pulse migrations, may be expressed as a succession of these 2-way split and admixture events.

We provide pseudocode to construct the event tree $\mathcal{T}$ in Algorithm 1. In words, we initially start with $\mathcal{K}$ equal to a collection of singleton sets corresponding to the leaves of $\mathcal{G}$. Processing each split or admixture event E back in time, we merge all blocks containing the child population(s) of E. Then we remove the child population(s) from this merged block, and add in the parent population(s). Within $\mathcal{T}$, the new merged block is the parent of the blocks removed at this stage.

We now describe Algorithm 2, the dynamic program to compute the joint SFS. We need a bit more notation. For population v, let n_v be the number of samples with ancestry in v; we will be keeping track of n_v lineages within population v. Let τ_v denote the amount of time between the top and bottom events of v, and for let $\eta_v(t)$ denote the scaled population size of v at time $t \in [0, \tau_v]$ above the base of v. Let $0 \le X_v^{(t)} \le n_v$ be the allele count within the n_v lineages of v at time t above its bottom, so $X_v^{(0)} = X_v$ in our earlier notation.

At event E in $\mathcal{T}$, let $K_E = \{v_1, \dots, v_{|K_E|}\}$ be the corresponding block of populations in $\mathcal{G}$. We define the conditional likelihood at E as

$$\ell_{\mathbf{x},\mathbf{z}}^{E,\mathbf{t}} = \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} = \mathbf{z} \mid \mathbf{X}_{K_E}^{(\mathbf{t})} = \mathbf{x}) \tag{4}$$

where $\mathbf{X}_{K_E}^{(\mathbf{t})} = (X_{v_1}^{(t_1)}, \dots, X_{v_{|K_E|}}^{t_{|K_E|}})$ is the vector of allele counts in populations K_E at times $\mathbf{t}$, and $\mathbf{X}_{\texttt{Leaves}(E)}$ is the observed data at the leaves beneath E.

In addition, we define the "partial SFS"

$$\phi_{\mathbf{z}}^{E} = \sum_{v \text{ descended from } K_E} \mathbb{E}[\text{branch length in } v \text{ with } \mathbf{X}_{\texttt{Leaves}(E)} = \mathbf{z}] \tag{5}$$

as the expected branch length at or below K_E subtending $\mathbf{z}$ leaves. Note that $\phi_{\mathbf{z}}^{\texttt{Root}(\mathcal{T})}$ gives the desired final result, and corresponds to equation (3) in the previous subsection.

For the remainder of the subsection, we will fix $\mathbf{X}_{\texttt{Leaves}(\mathcal{G})} = \mathbf{z}$, and drop the dependence on $\mathbf{z}$ in $\ell_{\mathbf{x},\mathbf{z}}^{E,\mathbf{t}}$ and $\phi_{\mathbf{z}}^{E}$. Algorithm 2 defines a dynamic program (DP) over the conditional likelihoods $\ell_{\mathbf{x}}^{E,\mathbf{0}}$ and partial SFS ϕ^E. The DP takes input vectors $\ell^1, \dots, \ell^D$ corresponding to the leaf populations $\text{Leaves}(\mathcal{G}) = \{1, \dots, D\}$. If the inputs $\ell^1, \dots, \ell^D$ are set to indicator vectors corresponding to the observed counts $X_1, \dots, X_D$, the DP of Algorithm 2 will return the corresponding SFS entry, as stated in the theorem below:

Theorem 1. *If* $(X_1, \dots, X_D) \neq \mathbf{0}, \mathbf{n}$*, then*

$$\phi = \text{DP}(\mathbf{e}_{X_1}, \dots, \mathbf{e}_{X_D}),$$

where $\mathbf{e}_{X_i} = (0, \dots, 1, \dots, 0) \in \mathbb{R}^{n_i+1}$ *denotes the vector with 1 at coordinate* X_i *and 0 elsewhere.*

We now present the formulas used by Algorithm 2, in a series of lemmas also used to prove Theorem 1. We start with a formula to "lift" $\mathbb{P}(\dots \mid \dots, X_v^{(0)}, \dots)$ up to $\mathbb{P}(\dots \mid \dots, X_v^{(\tau_v)}, \dots)$. That is, this formula transforms a likelihood conditioned on $X_v^{(0)}$, the allele count at the bottom of v, into a likelihood conditioned on $X_v^{(\tau_v)}$, the allele count at the top of v.

Algorithm 1 Construct event tree $\mathcal{T}$

procedure EVENTTREE($\mathcal{G}$)
  $\mathcal{K} \leftarrow \{\{l\} : l \in \text{Leaves}(\mathcal{G})\}$
  $V(\mathcal{T}) \leftarrow \{\}$ ▷ Vertices of $\mathcal{T}$
  $\mathcal{E}(\mathcal{T}) \leftarrow \{\}$ ▷ Edges of $\mathcal{T}$
  for events E in BackInTimeEvents($\mathcal{G}$) **do**
    if E is split **then**
      $w_1, w_2 \leftarrow$ ChildPopulations(E)
      $v \leftarrow$ ParentPopulation(E)
      $K_1 \leftarrow$ block in $\mathcal{K}$ containing w_1
      $K_2 \leftarrow$ block in $\mathcal{K}$ containing w_2
      $K_E \leftarrow K_1 \cup K_2 \setminus \{w_1, w_2\} \cup \{v\}$
      $\mathcal{K} \leftarrow \mathcal{K} \setminus \{K_1, K_2\} \cup \{K_E\}$
      $V(\mathcal{T}) \leftarrow V(\mathcal{T}) \cup \{K_E\}$
      if $K_1 \neq K_2$ **then**
        $\mathcal{E}(\mathcal{T}) \leftarrow \mathcal{E}(\mathcal{T}) \cup \{K_E \to K_1, K_E \to K_2\}$
      else
        $\mathcal{E}(\mathcal{T}) \leftarrow \mathcal{E}(\mathcal{T}) \cup \{K_E \to K_1\}$
      end if
    else if E is admixture **then**
      $w \leftarrow$ ChildPopulation(E)
      $v_1, v_2 \leftarrow$ ParentPopulations(E)
      $K' \leftarrow$ block in $\mathcal{K}$ containing w
      $K_E \leftarrow K' \setminus \{w\} \cup \{v_1, v_2\}$
      $\mathcal{K} \leftarrow \mathcal{K} \setminus \{K'\} \cup \{K_E\}$
      $V(\mathcal{T}) \leftarrow V(\mathcal{T}) \cup \{K_E\}$
      $\mathcal{E}(\mathcal{T}) \leftarrow \mathcal{E}(\mathcal{T}) \cup \{K_E \to K'\}$
    end if
  end for
  return $V(\mathcal{T}), \mathcal{E}(\mathcal{T})$
end procedure

Lemma 1 (Lifting)**.** *Let E be a split or admixture event with corresponding block K_E, and let $v \in K_E$ be a population within this block. Let $\mathbf{x}_{-v}$ a collection of allele counts on $K_E \setminus \{v\}$, and $\mathbf{t}_{-v}$ a collection of times on $K_E \setminus \{v\}$. Then, for $\mathbf{x} = k\mathbf{e}_v + \mathbf{x}_{-v}$ and $\mathbf{t} = \tau_v \mathbf{e}_v + \mathbf{t}_{-v}$, the conditional likelihood of E is*

$$\ell^{E,\mathbf{t}}_{\mathbf{x}} = \sum_{j=0}^{n_v} \left[e^{Q^{(n_v)} \int_0^{\tau_v} \frac{1}{\eta_v(t)} dt} \right]_{kj} \ell^{E,\mathbf{t}_{-v}}_{j\mathbf{e}_v + \mathbf{x}_{-v}}, \tag{6}$$

where $Q^{(n)} \in \mathbb{R}^{(n+1)\times(n+1)}$ is the transition rate matrix of the Moran model with n lineages; in particular, $Q^{(n)} = (q^{(n)}_{ij})_{0 \le i,j \le n}$ and

$$q^{(n)}_{ij} = \begin{cases} -i(n-i), & \text{if } i = j, \\ \frac{1}{2} i(n-i), & \text{if } |j - i| = 1, \\ 0, & \text{else.} \end{cases}$$

To process event E, we first use Lemma 1 to lift up the conditional likelihoods at the child populations, up to the time of E. We then apply one of Lemma 2, Lemma 3, or Lemma 4, to obtain the conditional likelihood

Algorithm 2 Dynamic program to compute the SFS ϕ

procedure DP($\ell^1, \ldots, \ell^D$) ▷ $\ell^i = \mathbf{e}_{X_i} \in \mathbb{R}^{n_i+1}$
  for event E in DepthFirstSearch($\mathcal{T}$) **do**
    if $K_E = \{d\}$ is leaf **then**
      $\ell^{E,\mathbf{0}} \leftarrow \ell^d$
    else if E is split event **then**
      $\ell^{E,\mathbf{0}} \leftarrow$ Lemmas 1 and 2
    else if E is join event **then**
      if |ChildEvents(E)| $= 1$ **then**
        $\ell^{E,\mathbf{0}} \leftarrow$ Lemmas 1 and 4
      else if |ChildEvents(E)| $= 2$ **then**
        $\ell^{E,\mathbf{0}} \leftarrow$ Lemmas 1 and 3
      end if
    end if ▷ Computed the conditional likelihood $\ell^{E,\mathbf{0}}$
    $\phi^E \leftarrow$ (10) ▷ Computes ϕ^E the partial SFS
  end for
  return $\phi^{\mathrm{Root}(\mathcal{T})}$ ▷ Return partial SFS at the root event
end procedure

at E from the lifted child likelihoods, depending on whether E is an admixture or split event, and whether the child populations of E fall into a single cluster or two independent clusters.

We first consider admixture events; Lemma 2 describes how to compute the conditional likelihood in this case. Let the child population be w and the parent populations be v_1, v_2. Each of the n_w lineages in w independently inherits from v_1 with probability q_1, or from v_2 with probability $q_2 = 1 - q_1$. So the number of lineages inheriting from v_1 is Binomial(n_w, q_1). Then, given that m_1 alleles are inherited from v_1 and $m_2 = n_w - m_1$ inherited from v_2, the particular alleles inherited from v_1 or v_2 are chosen by sampling without replacement.

Lemma 2. *(Admixture event) Let E be an admixture event, with child population w and parent populations v_1, v_2. Let E' be the child event in $\mathcal{T}$. Suppose each lineage in w comes from v_1 with probability q_1, and from v_2 with probability $q_2 = 1 - q_1$. For K_E the population cluster at E, let $\mathbf{x}_\cap$ be a vector of allele counts on $K_E \setminus \{v_1, v_2\}$. Then the conditional likelihood of allele counts $\mathbf{x}_\cap + x_1\mathbf{e}_{v_1} + x_2\mathbf{e}_{v_2}$ at E is given by*

$$\ell^{E,\mathbf{0}}_{\mathbf{x}_\cap + x_1\mathbf{e}_{v_1} + x_2\mathbf{e}_{v_2}} = \sum_{x_w=0}^{n_w} \ell^{E',\tau_w\mathbf{e}_w}_{\mathbf{x}_\cap + x_w\mathbf{e}_w} \sum_{\substack{m_1,m_2:\\ m_1+m_2=n_w}} \binom{n_w}{m_1} q_1^{m_1} q_2^{m_2} \sum_{\substack{j_1,j_2:\\ j_1+j_2=x_w}} \frac{\binom{x_1}{j_1}\binom{n_{v_1}-x_1}{m_1-j_1}}{\binom{n_{v_1}}{m_1}} \frac{\binom{x_2}{j_2}\binom{n_{v_2}-x_2}{m_2-j_2}}{\binom{n_{v_2}}{m_2}}. \tag{7}$$

We next consider a split event E, with parent population v and child populations w_1, w_2. We first consider the case where E has 2 distinct children in $\mathcal{T}$, i.e., w_1, w_2 fall into 2 distinct blocks beneath E. Denote the corresponding child events as E'_1, E'_2 respectively. Then the conditional likelihood at E is given by a convolution of the conditional likelihoods at E'_1, E'_2, as described in Lemma 3:

Lemma 3. *(Population split, 2 clusters) Let E be a split event with parent population v and child populations w_1, w_2. Suppose E has 2 child events E'_1, E'_2, with corresponding blocks $K_{E'_1}, K_{E'_2}$, where $w_1 \in K_{E'_1}, w_2 \in K_{E'_2}$. Let $\mathbf{x}_{-1}$ be allele counts on $K_{E'_1} \setminus \{w_1\}$ and $\mathbf{x}_{-2}$ be allele counts on $K_{E'_2} \setminus \{w_2\}$. Then the conditional*

likelihood at E is

$$\ell^{E,\mathbf{0}}_{\mathbf{x}_{-1}+\mathbf{x}_{-2}+x_v\mathbf{e}_v} = \sum_{\substack{x_1,x_2:\\ x_1+x_2=x_v}} \frac{\binom{n_{w_1}}{x_1}\binom{n_{w_2}}{x_2}}{\binom{n_v}{x_v}} \ell^{E_1',\tau_{w_1}\mathbf{e}_{w_1}}_{x_1\mathbf{e}_{w_1}+\mathbf{x}_{-1}} \ell^{E_2',\tau_{w_2}\mathbf{e}_{w_2}}_{x_2\mathbf{e}_{w_2}+\mathbf{x}_{-2}}. \tag{8}$$

Now consider the case where the population split E has just 1 child event E'. That is, the child populations w_1, w_2 fall into the same cluster beneath E. Then Lemma 4 describes how to obtain the conditional likelihood at E from the conditional likelihood at E'. This involves summing over the dimensions corresponding to w_1, w_2 within the conditional likelihood $\ell^{E',\tau_{w_1}\mathbf{e}_{w_1}+\tau_{w_2}\mathbf{e}_{w_2}}$. In addition, note that we may have $n_v < n_{w_1}+n_{w_2}$ (recall n_v is the number of samples with ancestry in v). That is, after merging w_1, w_2 backwards in time, we may be keeping track of more alleles than originally sampled, allowing us to "drop" some extraneous non-ancestral lineages, as illustrated in the root population of Figure 1. This is done by multiplying the pseudoinverse of a matrix B whose entries are hypergeometric probabilities.

Lemma 4. *(Population split, 1 cluster) Let E be a population split with exactly 1 child event E'. Denote the corresponding population clusters as K_E and $K_{E'}$. Denote the parent population as v, the child populations as w_1, w_2. Let $\mathbf{x}_\cap$ be a vector of allele counts on the populations in $K_E \cap K_{E'}$. Define $y^{\mathbf{x}_\cap} \in \mathbb{R}^{n_{w_1}+n_{w_2}+1}$ and $B \in \mathbb{R}^{(n_{w_1}+n_{w_2}+1)\times(n_v+1)}$ to be the 0-indexed arrays with entries*

$$y_i^{\mathbf{x}_\cap} = \sum_{\substack{j,k:\\ j+k=i}} \frac{\binom{n_{w_1}}{j}\binom{n_{w_2}}{k}}{\binom{n_{w_1}+n_{w_2}}{i}} \ell^{E',\tau_{w_1}\mathbf{e}_{w_1}+\tau_{w_2}\mathbf{e}_{w_2}}_{j\mathbf{e}_{w_1}+k\mathbf{e}_{w_2}+\mathbf{x}_\cap},$$

$$B_{i,j} = \frac{\binom{n_v}{j}\binom{n_{w_1}+n_{w_2}-n_v}{i-j}}{\binom{n_{w_1}+n_{w_2}}{i}}.$$

Then the conditional likelihood at E is given by

$$\ell^{E,\mathbf{0}}_{k\mathbf{e}_v+\mathbf{x}_\cap} = [B^+ y^{\mathbf{x}_\cap}]_k, \tag{9}$$

with B^+ denoting the Moore-Penrose pseudoinverse of B.

Finally, having computed the conditional likelihoods at event E, we wish to compute the partial SFS ϕ^E at the event. This is given by the recursive formula in Lemma 5, which involves the conditional likelihood at E, the expected number of mutations arising within each parent population v, and the partial SFS at the child events.

Lemma 5. *For an event E, let $K_E^{new} = K_E \setminus \left(\cup_{E'\prec E} K_{E'}\right)$ be the populations newly formed at E (i.e., formed by a population split or admixture at E). For $v \in K_E^{new}$, let $f_v(k)$ be the truncated SFS in population v,*

$$f_v(k) = \frac{1}{\theta}\mathbb{E}[\#\textit{mutations at } v \textit{ with } X_v = k],$$

which can be computed by the formulas in (Kamm et al., 2017). Then the partial SFS at E is given by

$$\begin{aligned}\phi^E = &\sum_{v\in K_E^{new}} \sum_{k=1}^{n_v} f_v(k)\ell^{E,\mathbf{0}}_{k\mathbf{e}_v} \\ &+ \begin{cases} 0, & \textit{if } E \textit{ is leaf event,} \\ \phi^{E'}, & \textit{if } \mathit{ChildEvents}(E) = \{E'\}, \\ \phi^{E_1'}\prod_{d\in \mathit{Leaves}(E_2')} \ell^d_0 + \phi^{E_2'}\prod_{d\in \mathit{Leaves}(E_1')} \ell^d_0, & \textit{if } \mathit{ChildEvents}(E) = \{E_1', E_2'\}. \end{cases}\end{aligned} \tag{10}$$

3.3 Normalizing constant and other linear functionals

To compute the probability $\frac{\phi_{\mathbf{z}}}{\|\phi\|_1}$ of a mutation having observed allele counts $\mathbf{z}$, we need not just $\phi_{\mathbf{z}}$, but also the normalizing constant $\|\phi\|_1 = \sum_{\mathbf{z}} \phi_{\mathbf{z}}$ the expected total branch length.

Computing $\|\phi\|_1$ directly is inefficient because of the $O(\prod_{d=1}^{D} n_d)$ possible entries $\mathbf{z}$. Instead, we can use Algorithm 2 to compute $\|\phi\|_1$, and many more statistics of the SFS, in the same time as a single entry:

Corollary 1. *For $\pi^d \in \mathbb{R}^{n_d+1}, d \in \{1, \dots, D\}$, the tensor dot product of the SFS ϕ against $\pi^1 \otimes \cdots \otimes \pi^D = [\pi^1_{z_1} \cdots \pi^D_{z_D}]_{z_1,\dots,z_D}$ is*

$$\begin{aligned}\phi \odot (\pi^1 \otimes \cdots \otimes \pi^D) &= \sum_{z_1,\dots,z_D} \phi_{z_1,\dots,z_D} \pi^1_{z_1} \cdots \pi^D_{z_D} \\ &= \mathrm{DP}(\pi^1, \dots, \pi^D) - \left(\prod_{d=1}^{D} \pi^d_0\right) \mathrm{DP}(\mathbf{e}_0, \dots, \mathbf{e}_0) - \left(\prod_{d=1}^{D} \pi^d_{n_d}\right) \mathrm{DP}(\mathbf{e}_{n_1}, \dots, \mathbf{e}_{n_D}).\end{aligned}$$

Corollary 1 says that for any rank-K tensor $A \in \mathbb{R}^{(n_1+1)\times\cdots\times(n_D+1)}$ with $A = \sum_{k=1}^{K} \mathbf{a}_1^{(k)} \otimes \cdots \otimes \mathbf{a}_D^{(k)}$,

$$\phi \odot A = \sum_{\mathbf{z}} \phi_{\mathbf{z}} A_{\mathbf{z}} = \sum_{k=1}^{K} \phi \odot (\mathbf{a}_1^{(k)} \otimes \cdots \otimes \mathbf{a}_D^{(k)})$$

can be computed in K calls to $\mathrm{DP}(\pi^1, \dots, \pi^D)$.[2] In particular, the expected total branch length $\|\phi\|_1$ is given by

$$\begin{aligned}\|\phi\|_1 = \sum_{\mathbf{z}} \phi_{\mathbf{z}} &= \phi \odot (\mathbf{1} \otimes \cdots \otimes \mathbf{1}) \\ &= \mathrm{DP}(\mathbf{1}, \dots, \mathbf{1}) - \mathrm{DP}(\mathbf{e}_0, \dots, \mathbf{e}_0) - \mathrm{DP}(\mathbf{e}_{n_1}, \dots, \mathbf{e}_{n_D})\end{aligned}$$

with $\mathbf{1}$ the vector with 1 at every coordinate.

Beyond the applications we explore here, we expect this result to be useful in a number of related settings. A number of population genetic statistics can be expressed as $\mathrm{f} \odot A$, including Watterson's estimator $\hat{\theta}_W$ of the mutation rate (Watterson, 1975), Fay and Wu's H statistic for positive selection (Fay and Wu, 2000), and Patterson's f_2, f_3, f_4 statistics for assessing topology (Patterson et al., 2012). Corollary 1 allows us to compute their expected values $\mathbb{E}[\mathrm{f} \odot A] = \theta\phi \odot A$, and to construct test statistics from the deviance $\mathrm{f} \odot A - \theta\phi \odot A$ under an appropriate null model.

An even wider class of population genetic statistics can be written as nonlinear functions of SFS-tensor products like $g(\frac{1}{\theta}\mathrm{f} \odot A_1, \frac{1}{\theta}\mathrm{f} \odot A_2, \dots)$; this class includes Tajima's D statistic for selection (Tajima, 1989), the F_{ST} statistic for population structure (Holsinger and Weir, 2009), and Patterson's D statistic for introgression (Patterson et al., 2012). These statistics may be viewed as plug-in estimators for $g(\phi \odot A_1, \phi \odot A_2, \dots)$, which we can compute with Corollary 1. Note that these estimators are biased due to the nonlinear function g, but the bias can be estimated via block jackknife, and will typically be small since $\frac{1}{\theta}\mathrm{f} \odot A \to \phi \odot A$ almost surely as the number of independent SNPs grows.

Another interesting linear statistic of the SFS that can be computed with Corollary 1 is $\mathbb{E}[T_{\mathrm{MRCA}}]$, the expected time of the most recent common ancestor. In particular, let d be any leaf population; for simplicity assume d is sampled at the present (i.e. d is not archaic). Then

$$\begin{aligned}\mathbb{E}[T_{\mathrm{MRCA}}] &= \sum_{z_1,\dots,z_D} \phi_{z_1,\dots,z_D} \frac{z_d}{n_d} \\ &= \phi \odot \left(\mathbf{1} \otimes \cdots \otimes \left(\frac{z_d}{n_d}\right)_{z_d \in \{0,\dots,n_d\}} \otimes \cdots \otimes \mathbf{1}\right).\end{aligned}$$

[2] The calls to $\mathrm{DP}(\mathbf{e}_0, \dots, \mathbf{e}_0)$ and $\mathrm{DP}(\mathbf{e}_{n_1}, \dots, \mathbf{e}_{n_D})$ only need to be computed once for all statistics.

To see this, note that T_{MRCA} is proportional to the expected number of mutations hitting an arbitrary lineage in d, and if a mutation has configuration $\mathbf{z} = (z_1, \ldots, z_D)$ derived copies, then the chance of hitting the lineage is $\frac{z_d}{n_d}$ (Zeng et al., 2006).

4 Application

We tested our method on a demographic inference problem in human genetics that is currently of interest. Lazaridis et al. (2014) showed that genetic variation in present-day Europeans suggests an admixture model involving three ancestral meta-populations: Ancient North Eurasian (ANE), Western Hunter Gatherers (WHG), and Early European Farmers (EEF). They also showed that EEF contains ancestry from a source that is an outgroup to all non-African populations, and yet shares much of the genetic drift common to non-African populations; they dubbed this ancestry component as "Basal Eurasian" ancestry. Later work (Lazaridis et al., 2016) showed that Basal Eurasian ancestry is shared by ancient and contemporary Middle Eastern populations, and is correlated with a decrease in Neanderthal ancestry, implying that Basal Eurasian ancestry contains lower levels of Neanderthal admixture when compared with non-Basal ancestry. The results from Lazaridis et al. (2014, 2016) were based on several related methods for modeling covariances in population allele frequencies, most notably qpGraph and qpAdm (Patterson et al., 2012; Haak et al., 2015). These methods are computationally efficient and robust, but are unable to infer the timing of demographic events.

We applied `momi2` to estimate the strength and timing of basal Eurasian admixture into early European farmers, and the split time of the basal Eurasian lineage. To do this, we built a demographic model relating 12 samples from 8 populations, shown in Figure 2. These samples consisted of the Altai Neanderthal (Prüfer et al., 2014); the 45,000 year old Ust'Ishim man from Siberia (Fu et al., 2014); 3 present-day populations (Mbuti, Sardinian, Han) with 3, 2, and 2 samples respectively; and 3 ancient samples representing the European ancestry components identified by Lazaridis et al. (2014): a 7,500 year old sample from the Linearbandkeramik (LBK) culture (representing EEF), an 8,000 year old sample from the Loschbour rock shelter in Luxembourg (representing WHG), and the 24,000 year old Mal'ta boy ("MA1") from Siberia (representing ANE). After data cleaning, our dataset consisted of 2.4×10^6 autosomal transversion SNPs. See Appendix A.3 for more details about the data.

To construct the topology of the model in Figure 2, we first obtained a tree by neighbor joining (Saitou and Nei, 1987), then added 3 extra admixture events reflecting prior knowledge, as well as a recent Neanderthal population decline starting at the Mbuti-Eurasian split. We inferred split times, population sizes (including the Neanderthal decline), and admixture times and proportions by maximizing a composite likelihood $\mathcal{L}$, given by the product of the likelihoods at every SNP:

$$\mathcal{L} = \prod_{s \in \text{SNPs}} \mathbb{P}(\mathbf{z}_s) \tag{11}$$

where $\mathbb{P}(\mathbf{z}_s) \propto \phi_{\mathbf{z}_s}$ and was computed by `momi2`. The low-coverage of the MA1 sample and the deep divergence of the Neanderthal sample may cause bias in SFS entries where only these samples contain derived alleles; we thus excluded these entries and corrected the normalizing constant appropriately (see Appendix A.3).

We used automatic differentiation to compute the gradient $\nabla \log \mathcal{L}$, which we used to search for the optimum of $\log \mathcal{L}$. We constructed nonparametric bootstrap confidence intervals by splitting the genome into 100 equally sized blocks, resampling these blocks to create 300 bootstrap datasets, and re-inferring the demography for each bootstrap dataset. We also used 300 parametric bootstraps to assess how well we could infer the demography under simulated data; for each parametric bootstrap dataset, we used `msprime` (Kelleher et al., 2016) to simulate ten 300 Mb chromosomes from our initial point estimate, and re-inferred the demography. Note the nonparametric bootstrap is better able to account for model misspecification, and we use it for all confidence intervals reported below.

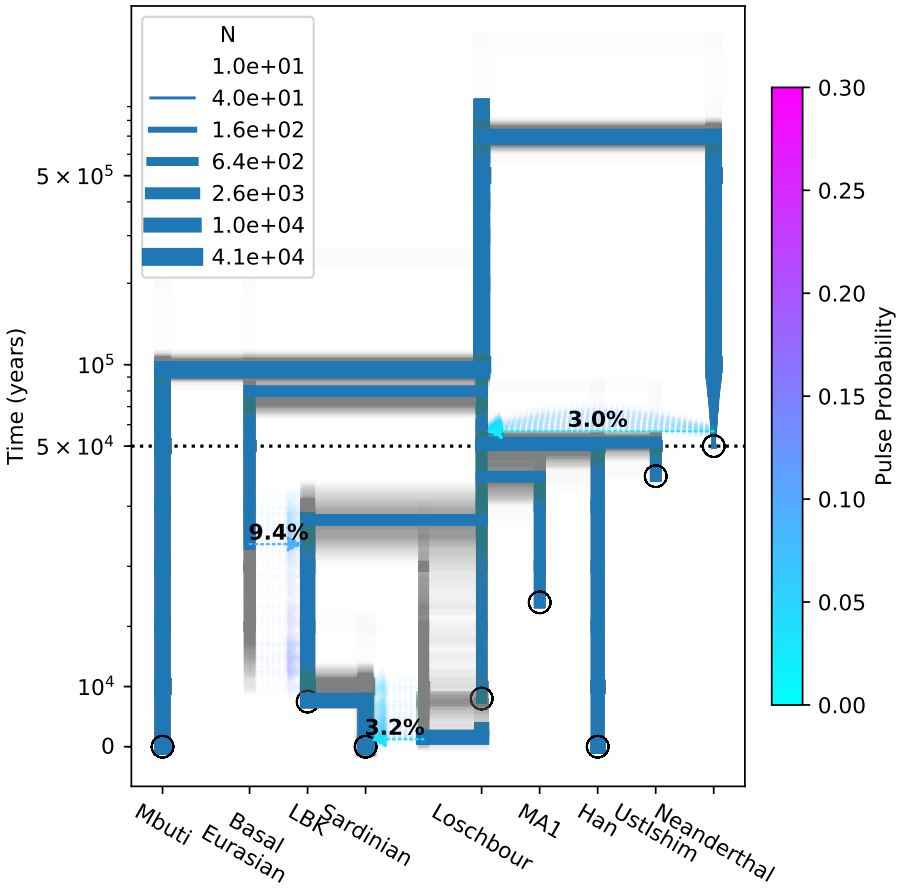


Figure 2: Inferred model and bootstraps for the 11 population demography described in Section 4. In the foreground (blue) is our point estimate from maximum composite likelihood; in the background (gray) are 300 bootstrap reestimates, which were created by splitting the data into 100 equally sized contiguous blocks, resampling these blocks with replacement, and refitting the model. The y-axis is linear below 5×10^4, then follows a logarithmic scale above 5×10^4.

Our inferred demography, along with nonparametric bootstrap re-estimates, are shown in Figure 2 and Table 2. Our parametric bootstrap estimates are shown in Figure 3. We inferred a pulse of 0.094 (95% CI of 0.049-0.174) from the ghost Basal Eurasian population to EEF ancestry (LBK), substantially less than the 0.44 inferred by (Lazaridis et al., 2014). This admixture was inferred to occur 33.7 kya (95% CI of 10.8-41.1 kya), shortly after the Loschbour-LBK split at 37.7 kya (95% CI of 32.2-42.3 kya). The split time of the ghost Basal Eurasian lineage from other Eurasians was inferred at 79.8 kya (95% CI of 67.4-101 kya). Other parameters were broadly in line with previous estimates, such as a Mbuti-Eurasian split of 96 kya, a Han-European split of 50 kya, a Neanderthal split of 696 kya, and Eurasians deriving 0.03 of their ancestry from Neanderthal (Terhorst et al., 2017; Green et al., 2010; Meyer et al., 2016).

Inferring the optimal demography from start to finish took 2.5 hours on a laptop with 4 CPU cores, and used 2 GB RAM. The 300 bootstraps were run separately on a high-performance compute cluster. To our knowledge, no other method can infer this demographic model using the full SFS. The `moments` software package (Jouganous et al., 2017) is capable of computing the SFS for up to 5 populations, less than the 8 populations here, though it can scale to more individuals per population than `momi2`. While the `fastsimcoal2`

Parameter	Estimate	Bias	SD	2.5%	97.5%
$N_{\texttt{Losch}}$	1.92e+03	14.1	185	1.57e+03	2.27e+03
$N_{\texttt{Mbu}}$	1.73e+04	−255	1.86e+03	1.6e+04	1.92e+04
$N_{\texttt{(Mbu,Losch)}}$	2.91e+04	−270	2.34e+03	2.7e+04	3.21e+04
$t_{\texttt{(Mbu,Losch)}}$	9.58e+04	−1.48e+03	9.32e+03	9.19e+04	1.03e+05
$N_{\texttt{Han}}$	6.3e+03	−17.1	400	5.71e+03	6.91e+03
$N_{\texttt{(Han,Losch)}}$	2.34e+03	−70.2	372	2.13e+03	2.7e+03
$t_{\texttt{(Han,Losch)}}$	5.04e+04	−171	2.75e+03	4.71e+04	5.43e+04
$t_{\texttt{(Ust,Losch)}}$	5.15e+04	−79.4	2.47e+03	4.85e+04	5.47e+04
$N_{\texttt{(Nean,Losch)}}$	1.82e+04	−216	1.59e+03	1.7e+04	1.99e+04
$t_{\texttt{(Nean,Losch)}}$	6.96e+05	−7.6e+03	5.74e+04	6.49e+05	7.58e+05
$t_{\texttt{Nean→Eurasian}}$	5.68e+04	−279	3.04e+03	5.38e+04	6.01e+04
$p_{\texttt{Nean→Eurasian}}$	0.0296	−2.82e−05	0.00252	0.0251	0.0349
$N_{\texttt{Nean}}$	86.9	−3.33	19.8	76.7	105
$t_{\texttt{(MA1,Losch)}}$	4.49e+04	98.2	2.36e+03	4.11e+04	4.87e+04
$N_{\texttt{LBK}}$	75.7	−685	629	4.12	1.9e+03
$t_{\texttt{(LBK,Losch)}}$	3.77e+04	543	2.59e+03	3.22e+04	4.23e+04
$p_{\texttt{Basal→EEF}}$	0.0936	−0.0122	0.0366	0.0485	0.174
$t_{\texttt{Basal→EEF}}$	3.37e+04	7.46e+03	1.01e+04	1.08e+04	4.11e+04
$t_{\texttt{(Basal,Losch)}}$	7.98e+04	−706	1.37e+04	6.74e+04	1.01e+05
$N_{\texttt{Sard}}$	1.5e+04	−1.28e+04	6.53e+04	8.58e+03	8.9e+04
$t_{\texttt{(Sard,LBK)}}$	7.69e+03	−1.69e+03	1.64e+03	7.51e+03	1.24e+04
$N_{\texttt{(Sard,LBK)}}$	1.2e+04	1.32e+03	1.99e+03	7.33e+03	1.45e+04
$t_{\texttt{GhostWHG→Sard}}$	1.23e+03	−2.93e+03	3.1e+03	597	1.06e+04
$p_{\texttt{GhostWHG→Sard}}$	0.0317	−0.00223	0.0151	0.00631	0.0618
$t_{\texttt{(GhostWHG,Losch)}}$	1.56e+03	−1.32e+04	1.07e+04	964	3.49e+04

Table 2: Estimated parameters of the demography in Figure 2, along with nonparametric bootstrap estimates of the bias and standard deviation, and 95% bootstrap quantiles. We use `(A,B)` to denote the ancestor of `A` and `B`; t_v and N_v to denote the height and size at vertex v; and $t_{\texttt{A→B}}$ and $p_{\texttt{A→B}}$ to denote respectively the time and strength of an admixture arrow from `A` to `B`.

software package (Excoffier et al., 2013) is capable of handling demographies of this size and larger, it doesn't compute the full, exact SFS, and also does not include an option for the ascertainment scheme we use here (excluding mutations private to Neanderthal and MA1).

We also used our model to produce estimates of the human mutation rate, which we estimated as 1.22×10^{-8} per base per generation, with a bootstrap-quantile 95% CI of $(1.12, 1.32) \times 10^{-8}$, closely matching previous estimates of the human mutation rate (Scally, 2016). To obtain this estimate, we compared the observed nucleotide diversity with that expected under the inferred demography, adjusting by the empirical transition to transversion ratio (Appendix A.3). Estimating the mutation rate was possible here because we did not use a prespecified mutation rate to estimate the model in Figure 2, instead using the known ages of the Ust'Ishim, LBK, Loschbour, and MA1 samples to calibrate dates.

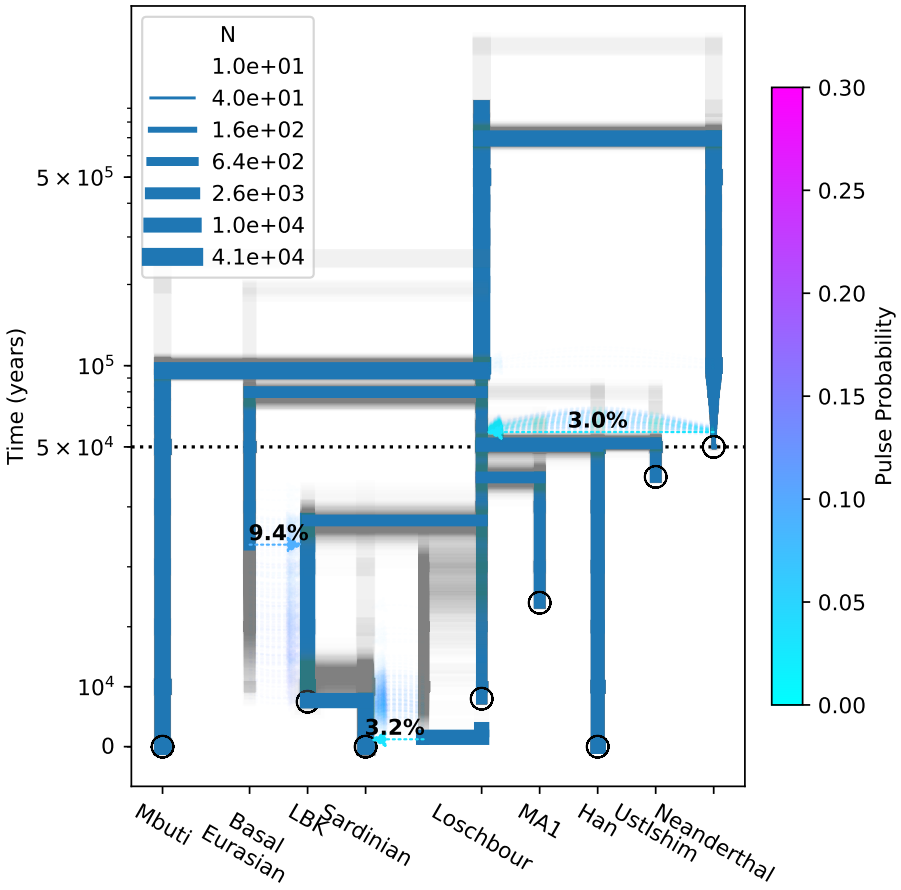


Figure 3: Parametric bootstraps for the demography inferred in Figure 2. The inferred demography (blue tree) was used to simulate 300 bootstrap replicates. Re-running our method on each replicate produced a new inferred demography which is plotted in gray.

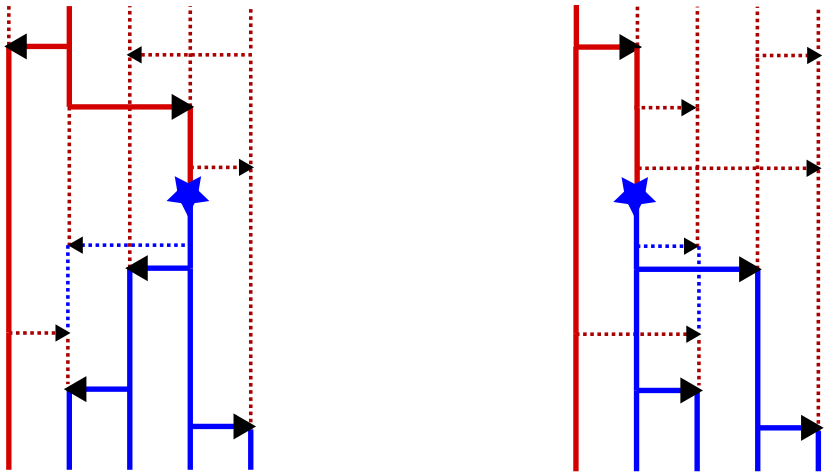
A Appendix

A.1 Model and notation

In this section, we formally define the stochastic process underlying our model, and introduce some additional notation needed for the proofs in Section A.2.

The main stochastic process we use is the *lookdown construction* of the Moran model (Donnelly and Kurtz, 1996; Donnelly et al., 1999), a variant of the Moran model where copying only occurs in one "direction" (as in Figure 4). We will make use of certain conditional independence properties that result from this one-way copying. However, we note that there is a simple coupling between the lookdown and standard Moran models, and these two models generate data with the same distribution.

We now describe our lookdown model in more detail. Within each of the D leaf populations we consider a countably infinite number of lineages, each with a unique label in $\mathbb{Z}_+$. We arbitrarily assign the $(n_1, \ldots, n_D)$ sampled lineages to the lowest labels $\{1, 2, \ldots, n_{\text{tot}}\}$, where $n_{\text{tot}} \equiv \sum_{d=1}^{D} n_d$, and arbitrarily assign the remaining unsampled lineages to labels $\{n_{\text{tot}} + 1, n_{\text{tot}} + 2, \ldots\}$. Each lineage extends infinitely backwards in time, and at each admixture event, each lineage randomly chooses the parent population it extends into. In addition, each lineage has an allele, which changes through time due to mutation and copying events.



(a) (b)

Figure 4: Standard and lookdown Moran models. (a) illustrates the standard Moran model, with copying in both directions; (b) illustrates the lookdown Moran model, with copying always from left to right. The sample genealogy (i.e., the lineages ancestral to present day samples) is in solid. The law of the genealogy is the coalescent under both models, and so the present-day samples of the two models are equal in distribution.

Similar to the usual Moran model, copying between a pair of lineages in population v occurs at rate $\frac{1}{\eta_v(t)}$, where $\eta_v(t)$ is the scaled population size of v. However, copying only occurs in one direction, from lineages with lower labels to lineages with higher labels (Figure 4). So if two lineages are labeled i and j respectively, with $i < j$, then the copying event $i \to j$ occurs at rate $\frac{1}{\eta_v(t)}$, whereas the reverse copying event $i \leftarrow j$ never happens (has rate 0). By contrast, in the standard (non-lookdown) Moran model, both $i \to j$ and $i \leftarrow j$ copying events happen at rate $\frac{1}{2\eta_v(t)}$. However, under both models, two lineages going backwards in time coalesce at rate $\frac{1}{\eta_v(t)}$, as in the coalescent. Thus, the sample genealogies under both models follow the same multipopulation coalescent distribution.

Denote the labeled lineages in population v, and their alleles at height t, by

$$\mathcal{M}_{v,t} = (\mathcal{M}_{v,t,(1)}, \mathcal{M}_{v,t,(2)}, \ldots)$$

where $\mathcal{M}_{v,t,(i)} = (\texttt{label}_{v,(i)}, \texttt{allele}_{v,t,(i)}) \in \mathbb{Z}_+ \times \{0,1\}$ is a pair, consisting of the ith lowest label at v, along with its corresponding allele at height $t \in [0, \tau_v)$. Note that we have ordered the lineages by their labels, so that $\texttt{label}_{v,(i)} < \texttt{label}_{v,(i+1)}$. Also note that we measure the time t from the bottom of v, so that $\texttt{allele}_{v,0,(i)}$ denotes the ith allele at the bottom of v, and $\texttt{allele}_{v,\tau_v,(i)}$ the ith allele at the top of v.

Let $X^{(t)}_{v,m} = \sum_{i=1}^{m} \mathbb{I}_{\{\mathcal{M}_{v,t,(i)} \text{ derived}\}}$ denote the number of derived alleles among the lowest m lineages at v, t. Let $n_v = \sum_{d \geq v} n_d$ be the number of samples in leaves below v. During computation, we will only need to keep track of the lowest n_v lineages in v, so we denote $X^{(t)}_v \equiv X^{(t)}_{v,n_v}$ to be the number of derived alleles among the first n_v lineages at v, t.

Finally, for an event E with $K_E = (v_1, \ldots, v_k)$ the corresponding block of populations in Algorithm 1, let $\mathbf{t} = (t_1, \ldots, t_k)$ be a corresponding set of times within each population $v_1, \ldots, v_k$. We define $\mathcal{M}_{E,\mathbf{t}} = (\mathcal{M}_{v_1,t_1}, \ldots, \mathcal{M}_{v_k,t_k})$ as the labeled alleles at each of $(v_1, t_1), \ldots, (v_k, t_k)$.

A.2 Proofs

The following two lemmas will be useful for several of the proofs below:

Lemma 6. *The distribution of $\mathcal{M}_{E,\mathbf{t}}$ is invariant to finite permutations of the labels within any population $v \in K_E$. Furthermore, the labels are independent of the alleles.*

Proof. By construction, none of the lineages within the populations in K_E are ancestral to each other. Thus the sample genealogy of any finite subsample of the lineages is the multipopulation coalescent, because going backwards in time, coalescence (copying) between each pair of lineages occurs at rate $\frac{1}{\eta_w(s)}$ at vertex w time s. The invariance to permutation, and independence of alleles and labels, follows from the coalescent. □

Lemma 7. *For event E and population $u \in K_E$, let $\mathcal{I} \subset \{n_u + 1, n_u + 2, \ldots\}$ be a (possibly random) collection of integers greater than n_u, and let $X_{u,\mathcal{I}}^{(t_u)}$ be the number of derived lineages in $\{\mathcal{M}_{u,t_u,(i)}\}_{i\in\mathcal{I}}$. Then $\mathbf{X}_{\mathtt{Leaves}(E)}$ is conditionally independent of $X_{u,\mathcal{I}}^{(t_j)}$ the allele counts on $\mathcal{I}$, given $\mathbf{X}_{K_E}^{(\mathbf{t})}$ the allele counts on the first $n_{v_1}, \ldots, n_{v_k}$ lineages of $K_E = \{v_1, \ldots, v_k\}$.*

Proof. Integrate over $\mathcal{M}_{E,\mathbf{t},\mathbf{n_v}} \equiv \{\mathcal{M}_{v,t_v,(i)}\}_{v\in K_E, 1\le i\le n_v}$ the first $(n_{v_1}, \ldots, n_{v_k})$ labeled alleles within K_E, and $\mathcal{M}_{u,t_u,\mathcal{I}} = \{\mathcal{M}_{u,t_u,(i)}\}_{i\in\mathcal{I}}$ the labeled alleles at $\mathcal{I}$:

$$\begin{aligned}
\mathbb{P}(\mathbf{X}_{\mathtt{Leaves}(E)} \mid \mathbf{X}_{K_E}^{(\mathbf{t})}, X_{u,\mathcal{I}}^{t_u}) &= \mathbb{E}[\mathbb{P}(\mathbf{X}_{\mathtt{Leaves}(E)} \mid \mathcal{M}_{E,\mathbf{t},\mathbf{n_v}}, \mathcal{M}_{u,t_u,\mathcal{I}}) \mid \mathbf{X}_{K_E}^{(\mathbf{t})}, X_{u,\mathcal{I}}^{(t_u)}] \\
&= \mathbb{E}[\mathbb{P}(\mathbf{X}_{\mathtt{Leaves}(E)} \mid \mathcal{M}_{E,\mathbf{t},\mathbf{n_v}}) \mid \mathbf{X}_{K_E}^{(\mathbf{t})}, X_{u,\mathcal{I}}^{(t_u)}] \\
&= \mathbb{E}[\mathbb{P}(\mathbf{X}_{\mathtt{Leaves}(E)} \mid \mathcal{M}_{E,\mathbf{t},\mathbf{n_v}}) \mid \mathbf{X}_{K_E}^{(\mathbf{t})}] \\
&= \mathbb{P}(\mathbf{X}_{\mathtt{Leaves}(E)} \mid \mathbf{X}_{K_E}^{(\mathbf{t})})
\end{aligned}$$

with the second equality because higher lineages cannot copy to lower lineages (so $\mathbf{X}_{\mathtt{Leaves}(E)} \perp \mathcal{M}_{u,t_u,\mathcal{I}} \mid \mathcal{M}_{E,\mathbf{t},\mathbf{n_v}}$), and the third equality because of the exchangeability and independence from Lemma 6 (so given $\mathbf{X}_{K_E}^{(\mathbf{t})}$, the alleles of $\mathcal{M}_{E,\mathbf{t},\mathbf{n_v}}$ are ordered by a uniform permutation independent of $X_{u,\mathcal{I}}^{(t_u)}$, and the labels of $\mathcal{M}_{E,\mathbf{t},\mathbf{n_v}}$ are independent of $\mathbf{X}_{K_E}^{(\mathbf{t})}, X_{u,\mathcal{I}}^{(t_u)}$). □

A.2.1 Proof of Theorem 1

Lemmas 2,4,3 give formulas for the partial likelihood $\ell_{\mathbf{x}}^{E,\mathbf{0}}$ using terms like $\ell_{\mathbf{x}'}^{E',\mathbf{t}'}$, where $E' \in \mathtt{Children}(E)$ and E is an admixture, 1-cluster split, or 2-cluster split, respectively. The terms $\ell_{\mathbf{x}'}^{E',\mathbf{t}'}$ in turn can be computed from the partial likelihoods $\ell_{\mathbf{x}'}^{E',\mathbf{0}}$ using Lemma 1; then Lemma 5 provides a formula for ϕ^E in terms of $\ell^{E,\mathbf{0}}$ and the partial SFS $\phi^{E'}$.

Algorithm 2 traverses the event tree $\mathcal{T}$ in a depth-first search, applying Lemmas 5,1,2,4,3 to compute $\ell_{\mathbf{x}}^{E,\mathbf{0}}, \phi^E$ at each event E from their values at the children of E. The input to Algorithm 2 are the likelihoods $\ell^{\{d\},\mathbf{0}}$ at the leaves, and the output is the partial SFS $\phi^{\mathtt{Root}(\mathcal{T})}$ at the root.

Thus, since

$$\ell^{\{d\},\mathbf{0}} = [\mathbb{P}(X_d \mid X_d = i)]_{0\le i\le n_d} = \mathbf{e}_{X_d},$$

it follows that $\mathrm{DP}(\mathbf{e}_{X_1}, \ldots, \mathbf{e}_{X_D}) = \phi^{\mathtt{Root}(\mathcal{T})} = \phi$.

A.2.2 Proof of Lemma 5

Without loss of generality assume $\theta = 1$. Then ϕ^E is the expected number of mutations at or below K_E with observed counts $\mathbf{X}_{\texttt{Leaves}(E)}$. We split ϕ^E into two parts: the mutations within $K_E^{\text{new}} = K_E \setminus \left(\cup_{E''' \prec E} K_{E'}\right)$ (i.e. the populations formed by a split or join at E); and the mutations that occur in $\cup_{E' \prec E} K_{E'}$, the populations that arise strictly below E.

The first part, of mutations at the new populations of E, is given by

$$\sum_{v \in K_E^{\text{new}}} \sum_{k=1}^{n_v} f_v(k) \ell^{E,\mathbf{0}}_{k\mathbf{e}_v},$$

since $f_v(k)$ is the expected number of mutations arising at v with $X_v = k$, and $\ell^{E,\mathbf{0}}_{k\mathbf{e}_v} = \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid X_v = k)$.

For the second part, of mutations strictly below E, we split into three cases: either E is a leaf event, or E has a single child event, or E has two child events. In the first case, if E is a leaf event, then no mutations can occur below E. In the second case, if E has a single child event E', then the expected number of mutations strictly below E is simply $\phi^{E'}$ by definition.

Finally, if E has two child events E'_1, E'_2, then E'_1 and E'_2 share no leaves, so a mutation underneath E'_1 will have no derived alleles in $\texttt{Leaves}(E'_2)$ and vice versa. Thus, the number of mutations strictly below E is

$$\phi^{E'_1} \prod_{d \in \texttt{Leaves}(E'_2)} \ell^d_0 + \phi^{E'_2} \prod_{d \in \texttt{Leaves}(E'_1)} \ell^d_0.$$

A.2.3 Proof of Lemma 1

Define a "quasi-lookdown" Moran model $\mathcal{M}^*$, which is identical to $\mathcal{M}$, except within the n_v lowest lineages of v, where we allow copying in both directions at rate $\frac{1}{2\eta_v(t)}$ (as in the non-lookdown Moran model).

For an event E with populations $K_E = (v_1, v_2, \ldots)$ and corresponding times $\mathbf{t} = (t_1, t_2, \ldots)$, let $\mathbf{X}^{(\mathbf{t})}_{K_E} = (X^{(t_1)}_{v_1}, X^{(t_2)}_{v_2}, \ldots)$ be the corresponding allele counts. Next, define $\mathbf{X}_{\preceq E,\mathbf{t}} = \{\mathbf{X}^{\mathbf{s}}_{K_{E'}}\}_{(E',\mathbf{s}) \preceq (E,\mathbf{t})}$ as the sample path of allele counts below $E, \mathbf{t}$, where $(E', \mathbf{s}) \preceq (E, \mathbf{t})$ if either E is above E', or $E = E'$ and $\mathbf{s} \leq \mathbf{t}$ component-wise. It will suffice to show $\mathbb{P}_{\mathcal{M}}(\mathbf{X}_{\preceq E,\mathbf{t}}) = \mathbb{P}_{\mathcal{M}^*}(\mathbf{X}_{\preceq E,\mathbf{t}})$, because then for $\mathbf{t} = \tau_v \mathbf{e}_v + \mathbf{t}_{-v}$,

$$\begin{aligned}
\ell^{E,\mathbf{t}}_{k\mathbf{e}_v + \mathbf{x}_{-v}} &= \sum_{j=0}^{n_v} \mathbb{P}_{\mathcal{M}}(\mathbf{X}^{(\mathbf{t}_{-v})}_{K_E} = j\mathbf{e}_v + \mathbf{x}_{-v} \mid \mathbf{X}^{(\mathbf{t})}_{K_E} = k\mathbf{e}_v + \mathbf{x}_{-v}) \ell^{E,\mathbf{t}_{-v}}_{j\mathbf{e}_v + \mathbf{x}_{-v}} \\
&= \sum_{j=0}^{n_v} \mathbb{P}_{\mathcal{M}^*}(\mathbf{X}^{(\mathbf{t}_{-v})}_{K_E} = j\mathbf{e}_v + \mathbf{x}_{-v} \mid \mathbf{X}^{(\mathbf{t})}_{K_E} = k\mathbf{e}_v + \mathbf{x}_{-v}) \ell^{E,\mathbf{t}_{-v}}_{j\mathbf{e}_v + \boldsymbol{\mu}_{-v}} \\
&= \sum_{j=0}^{n_v} \left[e^{M^{(n_v)} \int_0^{\tau_v} \frac{1}{\eta_v(t)} dt} \right]_{kj} \ell^{E,\mathbf{t}_{-v}}_{j\mathbf{e}_v + \boldsymbol{\mu}_{-v}}
\end{aligned}$$

as desired.

$\mathbb{P}_{\mathcal{M}}(\mathbf{X}_{\preceq E,\mathbf{t}}) = \mathbb{P}_{\mathcal{M}^*}(\mathbf{X}_{\preceq E,\mathbf{t}})$ follows from a coupling argument. Let $\mathcal{M}_{\preceq E,\mathbf{t}} = \{\mathcal{M}_{E',\mathbf{s}}\}_{(E',\mathbf{s}) \preceq (E,\mathbf{t})}$ the sample path of $\mathcal{M}$ below $E, \mathbf{t}$. We can map the partial sample paths of $\mathcal{M}^*_{\preceq \mathbf{v},\mathbf{t}}$ onto those of $\mathcal{M}_{\preceq \mathbf{v},\mathbf{t}}$ as follows: moving from the bottom to the top of v, whenever a lower label is copied over by a higher label, swap the labels of the lineages above the copying. Then the relabeled sample path has the same distribution as the lookdown construction, since the allele with the higher label is always copied over, and the rate of copying between pairs of lineages is $\frac{1}{\eta_v(t)}$. Since this relabeling also leaves $X^{(t)}_v$ unchanged, we have $\mathbb{P}_{\mathcal{M}}(\mathbf{X}_{\preceq E,\mathbf{t}}) = \mathbb{P}_{\mathcal{M}^*}(\mathbf{X}_{\preceq E,\mathbf{t}})$.

A.2.4 Proof of Lemma 2

First note that $\mathbf{X}_{\texttt{Leaves}(E)} = \mathbf{X}_{\texttt{Leaves}(E')}$ and

$$\begin{aligned}
\ell^{E,\mathbf{0}}_{\mathbf{x}_\cap + x_1\mathbf{e}_{v_1} + x_2\mathbf{e}_{v_2}} &= \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid \mathbf{X}^{(\mathbf{0})}_{K_E} = \mathbf{x}_\cap + x_1\mathbf{e}_{v_1} + x_2\mathbf{e}_{v_2}) \\
&= \sum_{x_w=0}^{n_w} \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid \mathbf{X}^{(\mathbf{0})}_{K_E} = \mathbf{x}_\cap + x_1\mathbf{e}_{v_1} + x_2\mathbf{e}_{v_2}, X^{(\tau_w)}_w = x_w)\, \mathbb{P}(X^{(\tau_w)}_w = x_w \mid \mathbf{X}^{(\mathbf{0})}_{K_E} = \mathbf{x}) \\
&= \sum_{x_w=0}^{n_w} \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid \mathbf{X}^{(\mathbf{0})}_{K_E} = \mathbf{x}_\cap + x_1\mathbf{e}_{v_1} + x_2\mathbf{e}_{v_2}, X^{(\tau_w)}_w = x_w) \\
&\qquad \times \sum_{\substack{m_1,m_2:\\ m_1+m_2=n_w}} \binom{n_w}{m_1} q_1^{m_1} q_2^{m_2} \sum_{\substack{j_1,j_2:\\ j_1+j_2=x_w}} \frac{\binom{x_1}{j_1}\binom{n_{v_1}-x_1}{m_1-j_1}}{\binom{n_{v_1}}{m_1}} \frac{\binom{x_2}{j_2}\binom{n_{v_2}-x_2}{m_2-j_2}}{\binom{n_{v_2}}{m_2}}
\end{aligned}$$

by sampling n_w alleles in w from n_{v_1}, n_{v_2} alleles in v_1, v_2, which are exchangeable by Lemma 6.

Next, consider the $n_{v_1} + n_{v_2} - n_w$ highest alleles of v_1, v_2, and let $\mathcal{I} \subset \{n_w + 1, n_w + 2, \dots\}$ be their relative positions within w. Then we conclude the proof by noting that $\mathbf{X}_{\texttt{Leaves}(E)} = \mathbf{X}_{\texttt{Leaves}(E')}$ and applying Lemma 7, to get

$$\begin{aligned}
&\mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid \mathbf{X}^{(\mathbf{0})}_{K_E} = \mathbf{x}_\cap + x_1\mathbf{e}_{v_1} + x_2\mathbf{e}_{v_2}, X^{(\tau_w)}_w = x_w) \\
&\quad = \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E')} \mid X_{v_1} = x_1, X_{v_2} = x_2, \mathbf{X}^{(\tau_w\mathbf{e}_w)}_{K'_E} = \mathbf{x}_\cap + x_w\mathbf{e}_w) \\
&\quad = \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E')} \mid X^{(\tau_w)}_{w,\mathcal{I}} = x_1 + x_2 - x_w, \mathbf{X}^{(\tau_w\mathbf{e}_w)}_{K'_E} = \mathbf{x}_\cap + x_w\mathbf{e}_w) \\
&\quad = \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E')} \mid \mathbf{X}^{(\tau_w\mathbf{e}_w)}_{K'_E} = \mathbf{x}_\cap + x_w\mathbf{e}_w) \\
&\quad = \ell^{E,\mathbf{0}}_{\mathbf{x}_c ap + x_w\mathbf{e}_w}.
\end{aligned}$$

A.2.5 Proof of Lemma 4

Let $\mathbf{X}_\cap = \mathbf{X}^{(\mathbf{0})}_{K_E} - X^{(0)}_v\mathbf{e}_v = \mathbf{X}^{(\mathbf{0})}_{K_{E'}} - X^{(0)}_{w_1}\mathbf{e}_{w_1} - X^{(0)}_{w_2}\mathbf{e}_{w_2}$ be the vector of allele counts on $K_E \cap K_{E'}$ at times $\mathbf{0}$. Then note that

$$\begin{aligned}
y_i^{\mathbf{x}_\cap} &= \sum_{\substack{j,k:\\ j+k=i}} \frac{\binom{n_{w_1}}{j}\binom{n_{w_2}}{k}}{\binom{n_{w_1}+n_{w_2}}{i}} \ell^{E',\tau_{w_1}\mathbf{e}_{w_1}+\tau_{w_2}\mathbf{e}_{w_2}}_{j\mathbf{e}_{w_1}+k\mathbf{e}_{w_2}+\mathbf{x}_\cap} \\
&= \sum_{\substack{j,k:\\ j+k=i}} \mathbb{P}(X^{(\tau_{w_1})}_{w_1} = j, X^{(\tau_{w_2})}_{w_2} = k \mid X^{(\tau_{w_1})}_{w_1} + X^{(\tau_{w_2})}_{w_2} = i, \mathbf{X}_\cap = \mathbf{x}_\cap)\, \ell^{E',\tau_{w_1}\mathbf{e}_{w_1}+\tau_{w_2}\mathbf{e}_{w_2}}_{j\mathbf{e}_{w_1}+k\mathbf{e}_{w_2}+\mathbf{x}_\cap} \\
&= \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid X^{(\tau_{w_1})}_{w_1} + X^{(\tau_{w_2})}_{w_2} = i, \mathbf{X}_\cap = \mathbf{x}_\cap)
\end{aligned}$$

with the second equality following from exchangeability (Lemma 6) and the third equality from $\mathbf{X}_{\texttt{Leaves}(E)} = \mathbf{X}_{\texttt{Leaves}(E')}$.

Next note that

$$\begin{aligned}
&\mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid X_{w_1}^{(\tau_{w_1})} + X_{w_2}^{(\tau_{w_2})} = i, \mathbf{X}_\cap = \mathbf{x}_\cap) \\
&\quad = \sum_{j=0}^{n_v} \mathbb{P}(\mathbf{X}_{K_E}^{(\mathbf{0})} = j\mathbf{e}_v + \mathbf{x}_\cap \mid X_{w_1}^{(\tau_{w_1})} + X_{w_2}^{(\tau_{w_2})} = i, \mathbf{X}_\cap = \mathbf{x}_\cap) \\
&\qquad\qquad \times \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid \mathbf{X}_{K_E}^{(\mathbf{0})} = j\mathbf{e}_v + \mathbf{x}_\cap, X_{w_1}^{(\tau_{w_1})} + X_{w_2}^{(\tau_{w_2})} = i) \\
&\quad = \sum_{j=0}^{n_v} \frac{\binom{n_v}{j}\binom{n_{w_1}+n_{w_2}-n_v}{i-j}}{\binom{n_{w_1}+n_{w_2}}{i}} \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid \mathbf{X}_{K_E}^{(\mathbf{0})} = j\mathbf{e}_v + \mathbf{x}_\cap, X_{w_1}^{(\tau_{w_1})} + X_{w_2}^{(\tau_{w_2})} = i)
\end{aligned}$$

with the second equality again due to exchangeability (Lemma 6).

Define $\mathcal{I} \subset \{n_v + 1, n_v + 2, \ldots\}$ so that $\{1, \ldots, n_v\} \cup \mathcal{I}$ are the indices in v of the first n_{w_1}, n_{w_2} alleles in w_1, w_2. Then because $X_{v,\mathcal{I}}^{(0)} = X_{w_1}^{(\tau_{w_1})} + X_{w_2}^{(\tau_{w_2})} - X_v^{(0)}$ and Lemma 7,

$$\mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid \mathbf{X}_{K_E}^{(\mathbf{0})}, X_{w_1}^{(\tau_{w_1})} + X_{w_2}^{(\tau_{w_2})}) = \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid \mathbf{X}_{K_E}^{(\mathbf{0})}, X_{v,\mathcal{I}}^{(0)}) = \mathbb{P}(\mathbf{X}_{\texttt{Leaves}(E)} \mid \mathbf{X}_{K_E}^{(\mathbf{0})})$$

and thus

$$y_i^{\mathbf{x}_\cap} = \sum_{j=0}^{n_v} \frac{\binom{n_v}{j}\binom{n_{w_1}+n_{w_2}-n_v}{i-j}}{\binom{n_{w_1}+n_{w_2}}{i}} \ell_{j\mathbf{e}_v+\mathbf{x}_\cap}^{E,\mathbf{0}} = \sum_{j=0}^{n_v} B_{ij} \ell_{j\mathbf{e}_v+\mathbf{x}_\cap}^{E,\mathbf{0}}$$

so letting $\boldsymbol{\ell}' = [\ell_{j\mathbf{e}_v+\mathbf{x}_\cap}^{E,\mathbf{0}}]_{0 \le j \le n_v} \in \mathbb{R}^{n_v+1}$, we have $y^{\mathbf{x}_\cap} = B\boldsymbol{\ell}'$ and therefore $\boldsymbol{\ell}' = B^+ y^{\mathbf{x}_\cap}$.

A.2.6 Proof of Lemma 3

Notice that

$$\begin{aligned}
\ell_{\mathbf{x}_{-1}+\mathbf{x}_{-2}+x_v\mathbf{e}_v}^{E,\mathbf{0}} &= \sum_{\substack{x_1,x_2: \\ x_1+x_2=x_v}} \mathbb{P}(\mathbf{X}_{K_{E_1'}}^{(\tau_{w_1}\mathbf{e}_{w_1})} = \mathbf{x}_{-1} + x_1\mathbf{e}_{w_1}, \mathbf{X}_{K_{E_2'}}^{(\tau_{w_2}\mathbf{e}_{w_2})} = \mathbf{x}_{-2} + x_2\mathbf{e}_{w_1} \mid \mathbf{X}_{K_E}^{(\mathbf{0})} = \mathbf{x}_{-1} + \mathbf{x}_{-2} + x_v\mathbf{e}_v) \\
&\qquad \times \ell_{x_1\mathbf{e}_{w_1}+\mathbf{x}_{-1}}^{E_1',\tau_{w_1}\mathbf{e}_{w_1}} \ell_{x_2\mathbf{e}_{w_2}+\mathbf{x}_{-2}}^{E_2',\tau_{w_2}\mathbf{e}_{w_2}} \\
&= \sum_{\substack{x_1,x_2: \\ x_1+x_2=x_v}} \frac{\binom{n_{w_1}}{x_1}\binom{n_{w_2}}{x_2}}{\binom{n_v}{x_v}} \ell_{x_1\mathbf{e}_{w_1}+\mathbf{x}_{-1}}^{E_1',\tau_{w_1}\mathbf{e}_{w_1}} \ell_{x_2\mathbf{e}_{w_2}+\mathbf{x}_{-2}}^{E_2',\tau_{w_2}\mathbf{e}_{w_2}}
\end{aligned}$$

with the first equality from the Markov property of the Moran process, and the second equality following from the exchangeability of the n_v alleles at vertex v (Lemma 6).

Population	Individuals	Alleles	kya
Mbuti	3	6	0
Han	2	4	0
Sardinian	2	4	0
Loschbour	1	2	7.5
LBK (Stuttgart)	1	2	8
MA1 (Mal'ta)	1	1	24
Ust'Ishim	1	2	45
Altai Neanderthal	1	2	50

Table 3: Populations and samples used for the example application in Section 4. We only used 1 random allele for MA1 due to low coverage. The ages of the samples are given in thousands of years ago (kya).

A.2.7 Proof of Corollary 1

Below, we will prove $\mathrm{DP}(\ell^1, \ldots, \ell^D)$ is a multilinear function of the input vectors $\ell^1, \ldots, \ell^D$. The result immediately follows from this, because then

$$\begin{aligned}
\phi \odot (\pi^1 \otimes \cdots \otimes \pi^D) &= \sum_{\mathbf{z} \neq \mathbf{0}, \mathbf{n}} \phi_{\mathbf{z}} \pi^1_{z_1} \cdots \pi^D_{z_D} \\
&= \sum_{\mathbf{z} \neq \mathbf{0}, \mathbf{n}} \mathrm{DP}(\pi^1_{z_1} \mathbf{e}_{z_1}, \ldots, \pi^D_{z_D} \mathbf{e}_{z_D}) \\
&= \mathrm{DP}(\sum_{z_1=0}^{n_1} \pi^1_{z_1} \mathbf{e}_{z_1}, \ldots, \sum_{z_D=0}^{n_D} \pi^D_{z_D} \mathbf{e}_{z_D}) - \mathrm{DP}(\pi^1_0 \mathbf{e}_0, \ldots, \pi^D_0 \mathbf{e}_0) - \mathrm{DP}(\pi^1_{n_1} \mathbf{e}_{n_1}, \ldots, \pi^D_{n_D} \mathbf{e}_{n_D}) \\
&= \mathrm{DP}(\pi^1, \ldots, \pi^D) - \mathrm{DP}(\pi^1_0 \mathbf{e}_0, \ldots, \pi^D_0 \mathbf{e}_0) - \mathrm{DP}(\pi^1_{n_1} \mathbf{e}_{n_1}, \ldots, \pi^D_{n_D} \mathbf{e}_{n_D}).
\end{aligned}$$

We now show $\mathrm{DP}(\ell^1, \ldots, \ell^D)$ is a multilinear function of $\ell^1, \ldots, \ell^D$. We start by showing that if event E has leaf populations $\mathtt{Leaves}(E) = (d_1, \ldots, d_L)$, then $\ell^{E,\mathbf{t}}$ is a multilinear functions of $\ell^{d_1}, \ldots, \ell^{d_L}$. We show this by induction over $(E, \mathbf{t})$. The base case, where $\mathbf{t} = \mathbf{0}$ and E is a leaf with $K_E = \{d_i\}$, is trivially true because $\ell^{E,\mathbf{0}} = \ell^{d_i}$.

For the next case, we note that (6), (7), and (9) express $\ell^{E,\mathbf{t}}$ as a tensor product of $\ell^{E',\mathbf{t}'}$ with a matrix, where $(E', \mathbf{t}') \prec (E, \mathbf{t})$ and $\mathtt{Leaves}(E) = \mathtt{Leaves}(E')$. $\ell^{E',\mathbf{t}'}$ is a multilinear function of $\ell^{d_1}, \ldots, \ell^{d_L}$ by induction, hence $\ell^{E,\mathbf{t}}$ is also.

Similarly, (8) expresses $\ell^{E,\mathbf{0}}$ as a product of $\ell^{E'_1,\mathbf{t}_1}$ and $\ell^{E'_2,\mathbf{t}_2}$, where $\ell^{E'_i,\mathbf{t}_i}$ is a multilinear function of $\{\ell^k : k \in \mathtt{Leaves}(E'_i)\}$ by induction, and furthermore $\mathtt{Leaves}(E'_1) \cap \mathtt{Leaves}(E'_2) = \emptyset$ and $\mathtt{Leaves}(E'_1) \cup \mathtt{Leaves}(E'_2) = \mathtt{Leaves}(E)$. Thus $\ell^{E,\mathbf{0}}$ is a multilinear function of $\ell^{d_1}, \ldots, \ell^{d_L}$.

Thus, each of the operations (6), (7), (8), (9) in Algorithm 2 preserves multilinearity of $\ell^{E,\mathbf{t}}$, and thus each $\ell^{E,\mathbf{t}}$ is a multilinear function of its leaf vectors by induction. Finally, a similar induction argument shows that each ϕ^E is a multilinear function of $\{\ell^k : k \in \mathtt{Leaves}(E)\}$, since (10) expresses ϕ^E as a sum of multilinear functions by induction.

A.3 Application supplement

A.3.1 Data

Table 3 gives the populations and samples we used for our example application in Section 4. All samples except for MA1 were taken from the SGDP dataset (Mallick et al., 2016). We added on the additional low-coverage MA1 sample (Raghavan et al., 2014) to represent the Ancient North Eurasian (ANE) component

of European ancestry. Due to the low-coverage of MA1, we only sampled a single random allele from it at each site, using a GATK pileup (McKenna et al., 2010) and restricting ourselves to reads with quality ≥ 30.

We ascertained SNPs at SGDP filter level 1, then used the genotype calls at filter level 0 at the ascertained SNPs. When ascertaining SNPs, we limited ourselves to sites that were polymorphic among the samples excluding MA1 and Neanderthal. We excluded MA1 during ascertainment due to its low coverage. We also excluded the Neanderthal sample during ascertainment because it had substantially fewer new mutations than expected based on its age; it was unclear whether this was due to changes in the mutation rate on that lineage since its deep split with modern humans, or whether this was an artifact of the SNP calling strategy used by SGDP. We used Corollary 1 to correct the normalizing constant of the SFS due to excluding MA1 and Neanderthal during ascertainment; in particular, we normalized the SFS by the total branch length of the subtree excluding MA1 and Neanderthal.

To avoid biases in ancient DNA caused by deamination (Dabney et al., 2013), we removed all transitions (i.e. A↔G and C↔T mutations), keeping only the transversions. We used Chimp as a proxy for the ancestral allele, removing all sites where the Chimp allele was missing. After data cleaning, we were left with 2,444,888 autosomal transversion SNPs that were segregating among the samples excluding MA1 and Neanderthal.

A.3.2 Model fitting procedure

We fit the model in Figure 2 in an iterative fashion, adding populations in one at a time and re-estimating the parameters. We started with a tree including the 4 populations Mbuti, Loschbour, Han, and Ust'Ishim, with no admixture events. This initial model had 8 parameters: the population sizes at the Han, Mbuti, and Loschbour leaves, the population size at the Eurasian and human ancestor (the Ust'Ishim leaf was set to the ancestral Eurasian population size), and the times that the Han, Ust'Ishim, and Mbuti populations diverged from Loschbour.

We next added in the Neanderthal population, with an admixture event from Neanderthal to the Eurasian ancestor. We added parameters for the Neanderthal-human split time, the Neanderthal-Eurasian introgression tie and strength, the ancestral Neanderthal-human population size, and a Neanderthal population exponential decline rate starting at the Eurasian-Mbuti split, following results from Prüfer et al. (2014).

We followed by adding on the MA1 sample, adding a single parameter for its split time (we fixed its population size to be the ancestral Eurasian population size). We then added the LBK early farmer, adding 4 parameters for its divergence time, Basal Eurasian admixture time and strength, and the split time of the Basal Eurasian lineage. Finally, we added in a Sardinian population, along with parameters for its population size, its split time, and admixture times and strength from the Loschbour WHG.

At each step, we re-estimated all parameters using the L-BFGS-B optimization algorithm, maximizing the multinomial composite likelihood (11). For the final few models (adding in LBK and then Sardinian) we initialized each estimation with a stochastic gradient descent before finishing with L-BFGS-B.

A.3.3 Mutation rate estimation

We used the within-population nucleotide diversity to estimate the mutation rate. The nucleotide diversity is the number of sites where 2 random alleles (drawn without replacement) differ. The empirical value of the nucleotide diversity in population i is

$$\hat{\pi}_i = \sum_{s \in \text{SNPs}} 2\frac{x_i^{(s)}(n_i - x_i^{(s)})}{n_i(n_i-1)} = \sum_{\mathbf{x}} 2\frac{x_i(n_i-x_i)}{n_i(n_i-1)}\mathrm{f}_{\mathbf{x}},$$

while the expected value of the nucleotide diversity per unit mutation rate is

$$\pi_i = \frac{1}{\theta}\mathbb{E}[\hat{\pi}_i] = \sum_{\mathbf{x}} 2\frac{x_i^{(s)}(n_i - x_i^{(s)})}{n_i(n_i-1)}\phi_{\mathbf{x}}.$$

This yields a mutation rate estimate $\hat{\theta}_i$ for each population, given by $\hat{\theta}_i = \frac{\hat{\pi}_i}{\pi_i}$. We then averaged over $\hat{\theta}_i$ to obtain our combined estimate $\hat{\theta} = \frac{1}{D}\sum_{i=1}^{D}\hat{\theta}_i$. To obtain the per-site mutation rate we need to divide by the number of bases L after our data cleaning process. We approximated this as follows: at SGDP filter level 1, there are about 2.13×10^9 sites per individual; multiply this by 0.93 due to excluding the sex chromosomes; multiply this by 0.32 due to only using transversions; finally, multiply by 0.93 to account for excluding sites missing the Chimp allele, and multiply by 1.1 to account for using genotype calls at filter level 0. The latter 2 factors (accounting for sites missing the Chimp allele and for adding in genotype calls at filter level 0) we estimated by observing how the number of heterozygotes per individual changed after these data cleaning steps.

Acknowledgments

This research is supported in part by an NIH grant R01-GM109454, a Packard Fellowship for Science and Engineering, the Human Frontiers Science Program LT000402/2017, and Wellcome Trust grants WT206194 and RG89781. YSS is a Chan Zuckerberg Biohub investigator.

References

Baharian, S. and S. Gravel (2018). On the decidability of population size histories from finite allele frequency spectra. *Theoretical Population Biology 120*, 42–51.

Beaumont, M. A. and R. A. Nichols (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences 263*(1377), 1619–1626.

Bhaskar, A. and Y. S. Song (2014). Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Annals of Statistics 42*(6), 2469–2493.

Bhaskar, A., Y. X. R. Wang, and Y. S. Song (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research 25*(2), 268–279.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez, K. E. Lohmueller, M. D. Adams, S. Schmidt, J. J. Sninsky, S. R. Sunyaev, et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics 4*(5), e1000083.

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution 29*(8), 1917–1932.

Chen, H. (2012). The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theoretical Population Biology 81*(2), 179–195.

Corliss, G., C. Faure, A. Griewank, L. Hascoet, and U. Naumann (2002). *Automatic Differentiation of Algorithms: From Simulation to Optimization*, Volume 1. New York: Springer Science & Business Media.

Coventry, A., L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell, J. Crosby, J. E. Hixson, T. J. Rea, D. M. Muzny, L. R. Lewis, et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications 1*, 131.

Dabney, J., M. Meyer, and S. Pääbo (2013). Ancient dna damage. *Cold Spring Harbor perspectives in biology 5*(7), a012567.

De Iorio, M. and R. C. Griffiths (2004). Importance sampling on coalescent histories. II: Subdivided population models. *Adv. Appl. Prob. 36*, 434–454.

De Maio, N., D. Schrempf, and C. Kosiol (2015). Pomo: An allele frequency-based approach for species tree estimation. *Systematic Biology 64*(6), 1018–1031.

Donnelly, P. and T. Kurtz (1996). A countable representation of the Fleming-Viot measure-valued diffusion. *Ann Probab 24*, 698–742.

Donnelly, P., T. G. Kurtz, et al. (1999). Particle representations for measure-valued population models. *The Annals of Probability 27*(1), 166–205.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution* (2nd ed.). Springer, New York.

Ewens, W. J. (2004). *Mathematical Population Genetics: I. Theoretical Introduction*. New York: Springer Science+Business Media, Inc.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics 9*(10), e1003905.

Excoffier, L. and M. Foll (2011). Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics 27*(9), 1332–1334.

Fay, J. C. and C. I. Wu (2000). Hitchhiking under positive darwinian selection. *Genetics 155*, 1405–1413.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution 17*, 368–376.

Fu, Q., H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. Johnson, A. Aximu-Petri, K. Prüfer, C. de Filippo, et al. (2014). Genome sequence of a 45,000-year-old modern human from western siberia. *Nature 514*(7523), 445.

Gazave, E., L. Ma, D. Chang, A. Coventry, F. Gao, D. Muzny, E. Boerwinkle, R. A. Gibbs, C. F. Sing, A. G. Clark, et al. (2014). Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences 111*(2), 757–762.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, C. D. Bustamante, D. L. Altshuler, et al. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences 108*(29), 11983–11988.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, et al. (2010). A draft sequence of the Neandertal genome. *Science 328*(5979), 710–722.

Griffiths, R. and S. Tavaré (1998). The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models 14*(1-2), 273–295.

Griffiths, R. C. and S. Tavaré (1997). Computational methods for the coalescent. In P. Donnelly and S. Tavaré (Eds.), *Progress in population genetics and human evolution*, Volume 87, pp. 165–182. Springer-Verlag, Berlin.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics 5*(10), e1000695.

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, et al. (2015). Massive migration from the steppe was a source for indo-european languages in europe. *Nature 522*, 207–211.

Holsinger, K. E. and B. S. Weir (2009). Genetics in geographically structured populations: defining, estimating and interpreting fst. *Nature Reviews Genetics 10*(9), 639–650.

Jenkins, P. A., J. W. Mueller, and Y. S. Song (2014). General triallelic frequency spectrum under demographic models with variable population size. *Genetics 196*(1), 295–311.

Jouganous, J., W. Long, A. P. Ragsdale, and S. Gravel (2017). Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics 206*(3), 1549–1567.

Kamm, J. A., J. Terhorst, and Y. S. Song (2017). Efficient computation of the joint sample frequency spectra for multiple populations. *Journal of Computational and Graphical Statistics 26*(1), 182–194.

Kelleher, J., A. M. Etheridge, and G. McVean (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology 12*(5), e1004842.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics 61*(4), 893–903.

Kingman, J. F. C. (1982). The coalescent. *Stoch. Process. Appl. 13*, 235–248.

Koller, D. and N. Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lauritzen, S. L. and D. J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological) 50*(2), 157–224.

Lazaridis, I., D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, S. Mallick, D. Fernandes, M. Novak, B. Gamarra, K. Sirak, et al. (2016). Genomic insights into the origin of farming in the ancient near east. *Nature 536*(7617), 419–424.

Lazaridis, I., N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, et al. (2014). Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature 513*(7518), 409–413.

Lukić, S. and J. Hey (2012). Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics 192*(2), 619–639.

Maclaurin, D., D. Duvenaud, and R. P. Adams (2015). Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*.

Mallick, S., H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, et al. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature 538*(7624), 201–206.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research 20*(9), 1297–1303.

Meyer, M., J.-L. Arsuaga, C. de Filippo, S. Nagel, A. Aximu-Petri, B. Nickel, I. Martínez, A. Gracia, J. M. B. de Castro, E. Carbonell, et al. (2016). Nuclear DNA sequences from the middle pleistocene sima de los huesos hominins. *Nature 531*(7595), 504–507.

Myers, S., C. Fefferman, and N. Patterson (2008). Can one learn history from the allelic spectrum? *Theor. Popul. Biol. 73*(3), 342–348.

Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. S. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science 337*(6090), 100–104.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics 154*(2), 931–942.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of mathematical biology 29*(1), 59–75.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich (2012). Ancient admixture in human history. *Genetics 192*(3), 1065–1093.

Pearl, J. (1982). Reverend bayes on inference engines: a distributed hierarchical approach. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 133–136.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. De Filippo, et al. (2014). The complete genome sequence of a neanderthal from the altai mountains. *Nature 505*(7481), 43–49.

Raghavan, M., P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen, I. Moltke, S. Rasmussen, T. W. Stafford Jr, L. Orlando, E. Metspalu, et al. (2014). Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature 505*(7481), 87–91.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution 4*(4), 406–425.

Sawyer, S. A. and D. L. Hartl (1992). Population genetics of polymorphism and divergence. *Genetics 132*(4), 1161–1176.

Scally, A. (2016). The mutation rate in human evolution and demographic inference. *Current opinion in genetics & development 41*, 36–43.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, W. J. Daly, and D. Altshuler (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res. 15*, 1576–1583.

Stephens, M. and P. Donnelly (2000). Inference in molecular population genetics. *J.R. Stat. Soc. Ser. B 62*, 605–655.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics 123*(3), 585–595.

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetics Research 52*(3), 213–222.

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

Terhorst, J. and Y. S. Song (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences 112*(25), 7677–7682.

Wakeley, J. and J. Hey (1997). Estimating ancestral population parameters. *Genetics 145*(3), 847–855.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology 7*(2), 256–276.

Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC bioinformatics 11*(1), 116.

Zeng, K., Y.-X. Fu, S. Shi, and C.-I. Wu (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics 174*(3), 1431–1439.