Fast and accurate approximation of the joint site frequency spectrum of multiple populations

Ethan M. Jewett

Corresponding author: Ethan M. Jewett (ejewett@gmail.com)

Keywords: site frequency spectrum, complex demography, approximation, coalescent.

Abstract

The site frequency spectrum (SFS) is a statistic that summarizes the distribution of derived allele frequencies in a sample of DNA sequences. The SFS provides useful information about genetic variation within and among populations and it can used to make population genetic inferences. Methods for computing the SFS based on the diffusion approximation are computationally efficient when computing all terms of the SFS simultaneously and they can handle complicated demographic scenarios. However, in practice it is sometimes only necessary to compute a subset of terms of the SFS, in which case coalescent-based methods can achieve greater computational efficiency. Here, we present simple and accurate approximate formulas for the expected joint SFS for multiple populations connected by migration. Compared with existing exact approaches, our approximate formulas greatly reduce the complexity of computing each entry of the SFS and have simple forms. The computational complexity of our method depends on the index of the entry to be computed, rather than on the sample size, and the accuracy of our approximation improves as the sample size increases.

1. Introduction

The site frequency spectrum (SFS) is a statistic that records the distribution of allele frequencies across one or more populations (Hartl and Clark 2007, Wakeley 2008). The distribution of allele frequencies contains information about the size history of a population and evolutionary factors such selective pressures (Watterson 1975, McCoy et al. 2014, Bhaskar and Song 2014, Tajima 1989, Fay and Wu 2000, Nielsen et al. 2005), making the SFS is a useful statistic for performing inference under population-genetic models (Marth et al. 2004, Keinan et al. 2007, Gutenkunst et al. 2009, Excoffier et al. 2013).

Many methods have been developed for computing the SFS among populations of time-varying size connected by migration. Methods based on the Wright-Fisher diffusion model (Gutenkunst et al. 2009, Gravel et al. 2011, Lukić and Hey 2012) are typically quite fast when the number of populations is small. However, for multiple populations, the SFS is a multidimensional array in which the number of dimensions equals the number of populations. Because diffusion-based methods must keep track of the full multidimensional distribution of allele frequencies when integrating forward in time, these approaches rapidly become computationally challenging as the number of populations increases.

In contrast to diffusion-based approaches, methods based on the coalescent model only consider the histories of alleles that are observed in the present-day sample. Thus, coalescent approaches allow the computation of the SFS term-by-term. Such approaches can lead to considerable improvements in both speed and accuracy for computing subsets of entries of the SFS, which can improve the efficiency of inference. Coalescent formulas for computing the expected SFS have been obtained for both single populations of time-varying size and for multiple populations connected by migration (Wakeley and Hey 1997, Chen 2012, Chen and Chen 2013). Recently, Kamm et al. (2017) greatly improved the efficiency and numerical stability of coalescent approaches by developing a recursive algorithm for computing the SFS in populations of time-varying size with pulse migrations among them. Jouganous et al. (2017) have also developed a method that is a variation on diffusion approaches, which allows the SFS to be computed for larger numbers of populations.

Although the methods of Kamm et al. (2017) and Jouganous et al. (2017) have made computation of the SFS extremely fast, it may be possible to reduce the complexity of computing terms of the SFS still further. A potential approach for reducing the complexity of coalescent methods is to derive accurate approximations of the SFS using deterministic approximations of the number of ancestral coalescent lineages remaining in the population at each time in the past (Griffiths 1984, Slatkin and Rannala 1997, Volz et al. 2009, Maruvka et al. 2011, Chen and Chen 2013, Jewett and Rosenberg 2014). These approximations can be used to derive accurate approximate formulas for a variety of useful population genetic quantities (Maruvka et al. 2011, Chen and Chen 2013, Jewett and Rosenberg 2014) and they can be computed under complicated demographic histories that are difficult for classical coalescent models (Jewett and Rosenberg 2014).

Chen and Chen (2013) derived an approximation of the single-population SFS using a formula by Polanski and Kimmel (2003), in which the single-population SFS is expressed as a sum over expected coalescent waiting times. The approximate formulas of Chen and Chen (2013), which were obtained by replacing exact expressions for expected coalescent waiting times in the Polanski and

Kimmel formula with accurate approximations, are fast and accurate for computing the SFS in a single population.

Here, we take an analogous approach to that of Chen and Chen (2013) to compute an accurate approximation of the SFS in a set of populations with migration. However, rather than using the expression from Polanski and Kimmel (2003) for the SFS as a sum over expected waiting times, we take a new approach, deriving formulas for the SFS under an approximate coalescent framework in which the number of lineages as a function of time is deterministically equal to its expected value. Using this approach, we show that the SFS can be expressed as a sum over the expected total length of branches ancestral to subsets of sampled sequences. As we will see, this alternative formulation of the SFS provides a simple and intuitive way of deriving approximations of the SFS in multiple populations connected by migration. This approach allows us to obtain formulas for the SFS that have simple expressions and low computational complexity.

2. A summary of the main results

In a single population, the SFS for a sample of n sequences is an ordered tuple $\boldsymbol{\xi}_n = (\xi_{n,1}, \ldots, \xi_{n,n-1})$ of length $n-1$ in which the ith term records the number of polymorphic sites at which the derived allele appears in exactly i out of n sequences. For P populations with n_p sequences sampled from population p, the SFS is a P-dimensional array in which entry $i_1, \ldots, i_P$ records the number of polymorphic sites at which the derived allele appears in i_p copies in population p, for $p = 1, \ldots, P$. We first present approximate formulas for the SFS in the case of a single population of piecewise constant size and then consider more complicated models. All derivations are deferred to Section 3.

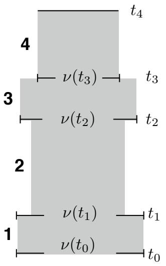
2.1. Computing the SFS in a single population.

Consider a single population of piecewise constant size like the one shown in the figure in Box 1. In such a population, the size $N(t)$ at time t changes over the course of K different time intervals $\{t_{k-1}, t_k\}_{k=1}^{K}$ satisfying $t_0 < t_1 < \cdots < t_K \leq \infty$ in such a way that $N(t) = N(t_{k-1})$ for $t \in [t_{k-1}, t_k)$. We denote the relative population size at time t by $\nu(t) = N(t)/N$ for some reference effective population size N, where time is measured in units of $2N$ generations.

In Section 3.3, we show that our approach yields the exact formula for the SFS given in Equation (1), which is equivalent to Equation (10) of Kamm et al. (2017). Because the formula appears in a general form expressed as an integral in Kamm et al. (2017), we report the specific simple form for a piecewise constant population here for completeness. Note that when $t_K < \infty$, Equation (1) computes the truncated SFS derived in Kamm et al. (2017). In Section 3.4, we also show that the expectation of $\boldsymbol{\xi}_n$ in a population of piecewise constant size can be approximated using Equation (3) in Box 1.

The approximate formula for the SFS in Box 1 provides a fast method for computing the SFS that is simple to implement. Although diffusion approaches are still the fastest methods for computing the complete SFS for large n, Equation (3) provides an improvement in efficiency over existing methods for computing subsets of entries. In particular, the complexity of Equation (3) does not depend on the sample size n. Instead, it depends linearly on the index i of the computed term. The complexity for computing the first L terms of the SFS is $O(L^2K)$, where K is the number of

Box 1 Computing $\mathbb{E}\boldsymbol{\xi}_n$ in a single population of piecewise constant size.



Consider a population size history defined piecewise over K different time intervals $\{[t_{k-1}, t_k)\}_{k=1}^{K}$ for times satisfying $t_0 < t_1 < \cdots < t_K \leq \infty$, where time is measured in coalescent units of $2N$ generations with respect to a reference population of arbitrary size N diploids. Suppose that the relative population size in the kth time interval is given by $\nu(t) = \nu(t_{k-1})$ for $t \in [t_{k-1}, t_k)$. An example of such a population history with four time intervals is shown on the left.

The exact expectation of the ith entry of the classical SFS $\boldsymbol{\xi}_n$ is

$$\mathbb{E}\xi_{n,i} = \frac{\theta}{2}\sum_{k=1}^{K}\nu(t_{k-1})\sum_{m=2}^{n} W_{n,i,m}\, e^{-\binom{m}{2}\tau_{k-1}}\binom{m}{2}^{-1}\left[1 - e^{-\binom{m}{2}\frac{t_k - t_{k-1}}{\nu(t_{k-1})}}\right] \qquad (1)$$

where

$$W_{n,i,m} = -(2m-1)\binom{n}{i}\sum_{j=0}^{i}(-1)^{i-j}\binom{i}{j}\frac{(n-j)_{[m]}}{(n-j)_{(m)}}. \qquad (2)$$

The quantity $W_{n,i,m}$ can also be computed efficiently using the following recursion from Kamm et al. (2017)

$$W_{n,i,m} = \frac{6}{n+1}, \text{ if } m = 2$$

$$W_{n,i,m} = 30\frac{n-2i}{(n+1)(n+2)}, \text{ if } m = 3$$

$$W_{n,i,m} = -\frac{(m-1)(2m-1)(n-m+2)}{(m-2)(2m-5)(n+m-1)}W_{n,i,m-2} + \frac{(2m-1)(n-2i)}{(m-2)(n+m-1)}W_{n,i,m-1}, \text{ if } m > 3.$$

An approximate formula: The expectation of the ith term of the SFS can be approximated by

$$\mathbb{E}\xi_{n,i} \approx \binom{n}{i}\theta\sum_{k=1}^{K}\nu(t_{k-1})\sum_{j=0}^{i}(-1)^{i-j+1}\binom{i}{j}\log\left[1 + \frac{(n-j)[e^{(t_k - t_{k-1})/2\nu(t_{k-1})} - 1]}{n-j-(n-j-1)e^{-\tau_{k-1}/2}}\right] \qquad (3)$$

whenever $t_K < \infty$, where $\tau_k = \sum_{j=1}^{k}(t_j - t_{j-1})/\nu(t_{j-1})$ and $\frac{\theta}{2}$ is the mutation rate in the coalescent.

piecewise constant epochs, which is lower than the $O(LnK)$ complexity of Equation (1), which is the current state of the art.

Panels A and E of Figure 1 show the runtimes of evaluating the formulas in Box 1 implemented in Mathematica, compared with those of the SFS software packages *momi* and $\partial a \partial i$. For all plots in Figure 1, the SFS was computed for a population with a bottleneck. The population size history is given by $\nu(t_0) = 1$, $\nu(t_1) = 0.5$, $\nu(t_2) = 1$ at times $t_0 = 0$, $t_1 = 0.1$, $t_2 = 0.2$, and $t_3 = \infty$. When computing Equation (3), we require $t_3 < \infty$ so we chose the large value $t_3 = 20$. As expected, the runtime of the approximate formula (Equation 3) is constant in n, whereas the runtimes of *momi* and $\partial a \partial i$ increase with n.

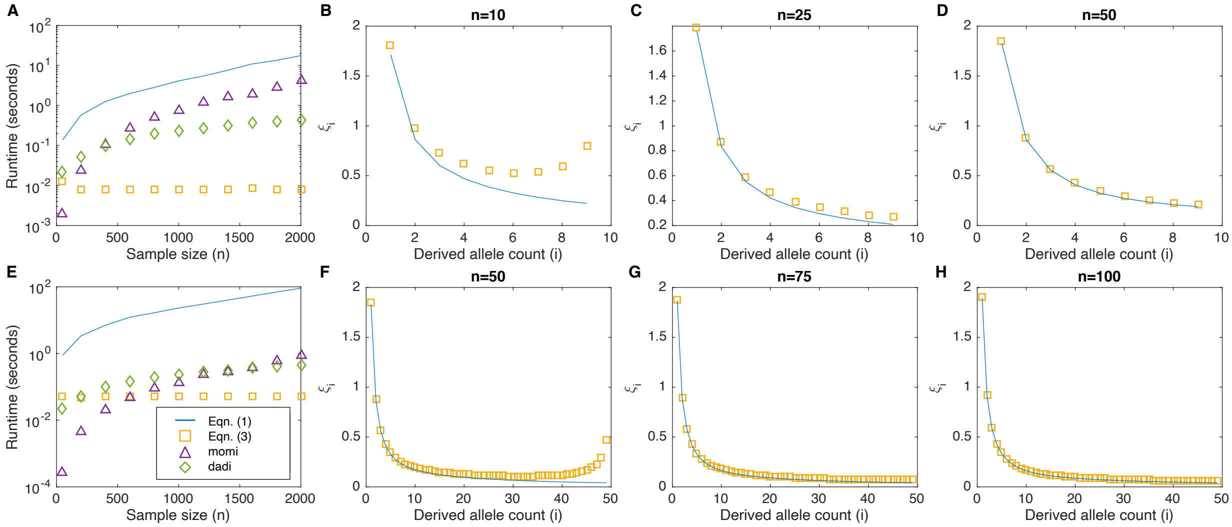


Figure 1. Accuracy and timing of the formulas in Box 1 for a population with a bottleneck. All panels correspond to a population history in which the size is given by $\nu(t_0) = \nu(t_2) = 1$, $\nu(t_1) = 0.5$, at times $t_0 = 0$, $t_1 = 0.1$, and $t_3 = 0.2$. Panel A shows the runtime for computing the top 9 entries of the SFS for different sample sizes. Panels B-D show the accuracy in the top 9 entries of the SFS as the sample size n is increased from 10 to 50. Panel E shows the runtime for computing the top 49 entries, and panels F-H show the accuracy in the top 49 entries as the sample size n is increased from 50 to 100.

It is evident from Panels A and E of Figure 1 that, even though the asymptotic complexity of Equation (1) is the same as that of *momi*, additional speed-ups obtained in the implementation of *momi* make *momi* extremely fast for computing low-order terms of the SFS when sample sizes are small to moderate. In fact, both *momi* and $\partial a \partial i$ are fast in the case of a single population.

2.2. **Computing the joint SFS in multiple populations with pulse migrations.** For samples taken from P different populations, the SFS is a P-dimensional array. In particular, if n_p homologous sequences are sampled from population p ($p = 1, \dots, P$), then $\boldsymbol{\xi}$ is an array with dimensions $n_1 \times n_2 \times \cdots \times n_P$ in which entry $(i_1, \dots, i_P)$ records the number of polymorphic sites at which the derived allele appears on i_1 sequences from the first population, i_2 sequences from the second population, and so on.

We show in Section 3 that the SFS for a collection of populations of piecewise constant size connected by instantaneous pulse migrations can be approximated using Equation (4) in Box 2. Figure 3 shows the accuracy and timing of Equation (4) compared with *momi* and $\partial a \partial i$ for the case of two populations of sizes $\nu_1 = 3$ and $\nu_2 = 2$ that split from an ancestral population of size $\nu_A = 1$ at $t = 0.05$ coalescent time units in the past. From Panels A-F of Figure 3, it can be seen that the approximate formula accurately captures the lower-order terms of the SFS even when the number of samples is small (e.g., $n_1 = n_2 \approx 20$).

There is always a certain amount of error in the higher order terms of the SFS, which is visible as the right-hand-side of Panels A-C of Figure 3. However, for lower order terms of the SFS, the approximation becomes more accurate than $\partial a \partial i$ as the sample size increases. This can be seen in panels G, H and I of Figure 3, which compare errors in the approximation with errors in $\partial a \partial i$, taking

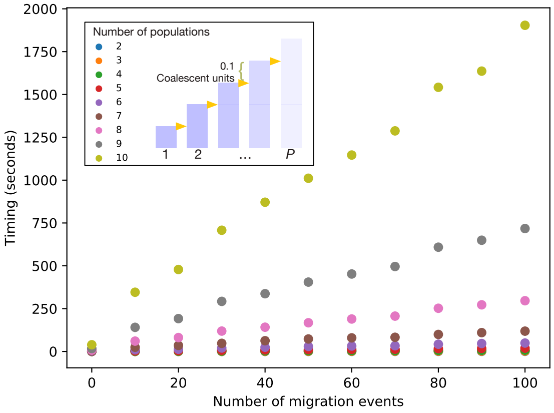


FIGURE 2. Time for computing the first term of the SFS using the approximation in Box 2, for different numbers of populations and pulse migration events. The form of the population history is shown in the inset diagram. In this model, a new population branches from its ancestral population every 0.1 coalescent units and each population has constant size $N = 10,000$. Migrations from the descendant population to the ancestral population are evenly spaced between the present-day and the founding event and 10% of the lineages in the descendant population migrate in each event. The sample size from each population is 50 lineages.

the values of *momi* as the truth. Panels G and H show average absolute errors, $\xi_{i,j}^{Approx} - \xi_{i,j}^{momi}$ and $\xi_{i,j}^{\partial a \partial i} - \xi_{i,j}^{momi}$, whereas Panel I shows average percent errors, $(\xi_{i,j}^{Approx} - \xi_{i,j}^{momi})/\xi_{i,j}^{momi}$ and $(\xi_{i,j}^{\partial a \partial i} - \xi_{i,j}^{momi})/\xi_{i,j}^{momi}$.

As expected, the absolute error in the approximation is greatest in the terms of greatest magnitude, which are typically the low order terms of the SFS. This can be seen in Figure 3G, which shows the mean absolute difference between the approximation and *momi* for different subsets of SFS terms. In particular, the yellow solid line shows the average absolute error across the first 5×5 terms of the SFS, whereas the dashed dotted yellow line shows the average absolute error across the expanded set containing the first 20×20 terms. Because the solid line in Figure 3G is above the dashed and dashed-dotted lines, it can be seen that the average absolute error is greater for the first 5×5 terms than for the higher order terms.

The fact that the low-order (high magnitude) terms of the SFS have the greatest absolute error is supported by figure Figure 3H, which shows that the maximum absolute error is the same when considering the first 5×5 terms, the first 10×10 terms, or the first 20×20 terms (the solid, dashed, and dashed-dotted lines are on top of one another), indicating that the maximum absolute error is in the set of first 5×5 term. In contrast, the mean relative error increases as we consider higher order terms of the SFS (Figure 3I), suggesting that the greatest relative error is in the higher order terms.

In Figures 3G-H, we have also compared $\partial a \partial i$ to *momi* (green lines). From Figures 3G-H it can be seen that the diffusion approximation in $\partial a \partial i$ diverges from the values computed by *momi* as the sample size increases.

When computing the low-order entries of the SFS in the case of multiple populations, the approximate SFS is also faster than existing approaches. Figure 3K shows the computation time of the first 11×11 entries of the SFS corresponding to counts counts $i_1 = 0, ..., 10$ in population 1 and $i_2 = 0, ..., 10$ in population 2. From Figure 3K, it can be seen that the approximation can be computed in constant time in the sample size n, making it faster than the current state of-the-art approaches for computing terms of the SFS when the sample size is large. Although the approximation is the fastest approach for computing each individual SFS term, Figure 3L suggests that $\partial a \partial i$ is still the fastest method for computing the full spectrum.

Figure 2 shows the computation time for the approximation under increasingly complicated demographic histories with increasing numbers of populations and pulse migration events. In particular, we modeled multiple populations by considering a serial founder model in which each population gave rise to an offspring population via a pulse migration consisting of 10% of its lineages. To evaluate the computation time of the approximation for p populations, we considered the youngest p populations. To evaluate the computation time in the presence of pulse migrations, we we added increasing numbers of pulse migrations between each population and its parent population, evenly spaced across the age of the younger population. The computation time grows quadratically in the number of populations and linearly in the number of pulse migration events (Figure 2).

3. Derivation of the formulas

In this section we derive the equations presented in Sections 2.1 and 2.2. Our approach is to first express the SFS as a sum of expected total branch lengths ancestral to different subsets of sampled sequences. This approach is analogous to that of Polanski and Kimmel (2003) who expressed the SFS as a sum of expected coalescence waiting times. As we will see, the exact formulas for the SFS obtained by the two approaches are the same; however, our approach of integrating over branch lengths makes it straightforward to derive fast approximate formulas in the case of multiple populations.

To facilitate the derivations, we define notation that we will use throughout this section. For a set of n sequences sampled from a single population, let S_i denote a particular subsample of i sequences. Let $\pi(S_i)$ denote the number of sites at which the derived allele is private to the sample S_i and let $\alpha(S_i)$ denote the number of sites at which the derived allele is ancestral to all of the i sequences in the sample, and to no others. The quantities $\pi(S_i)$ and $\alpha(S_i)$ are illustrated in Figure 4 for a subsample of size $i = 3$ sequences at 3 different loci.

The quantity $\alpha(S_i)$ is closely related to the ith term of the $\boldsymbol{\xi}$, which is simply the sum of $\alpha(S_i)$ over all subsets of size i; the summation yielding the total number of segregating sites that appear in exactly i lineages. To compute $\alpha(S_i)$ directly, we could consider the number of lineages that are ancestral to all members of S_i at each time t in the past and then integrate this quantity over all time, multiplied by the rate of new mutations. Our approach is to observe that the quantity $\pi(S_i)$ is considerably easier to compute, being the integral over the total number of branches ancestral to the set S_i. The quantity $\alpha(S_i)$ is related to $\pi(S_i)$ via the principle of inclusion and exclusion. We now provide the details of this computation.

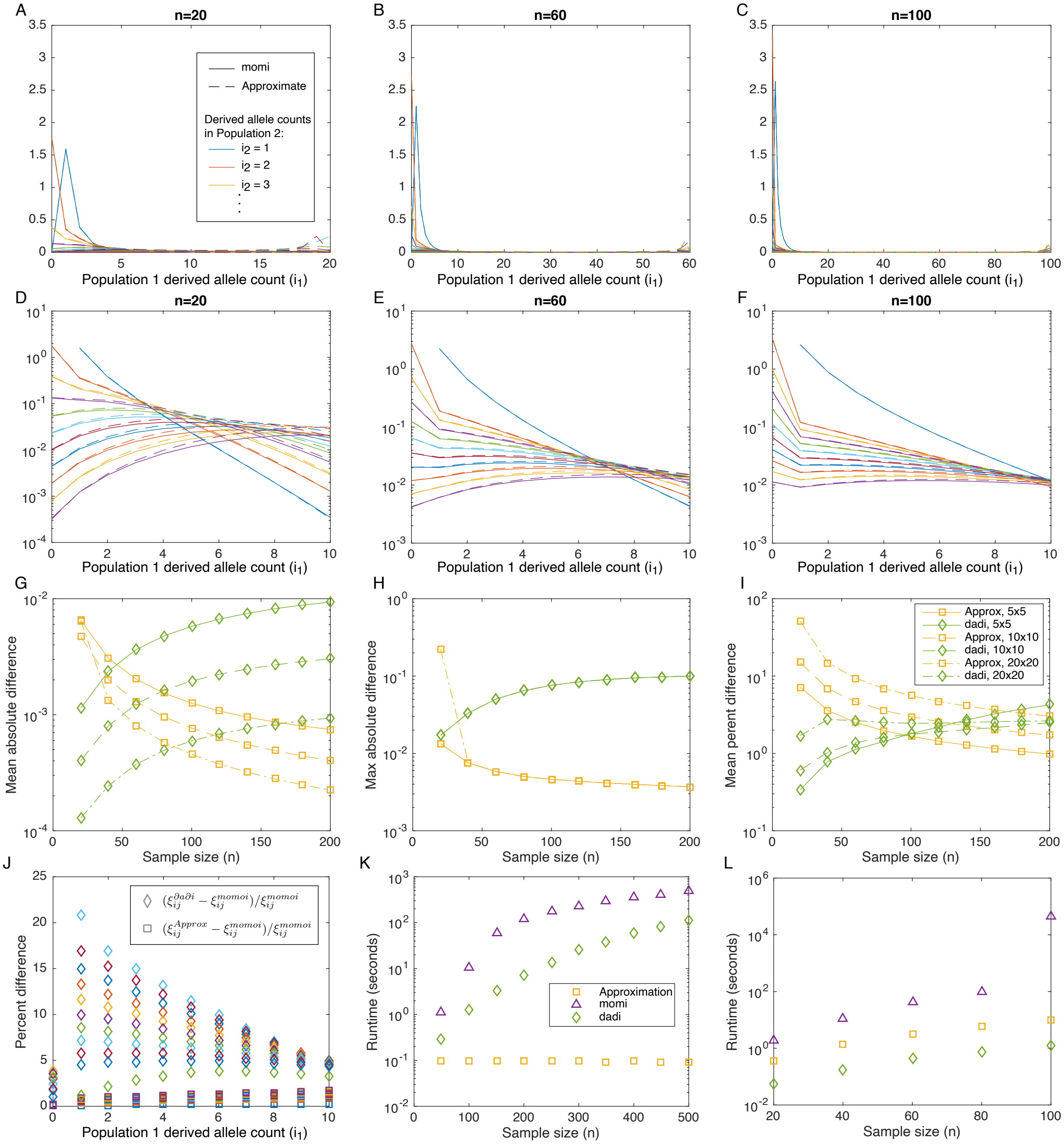
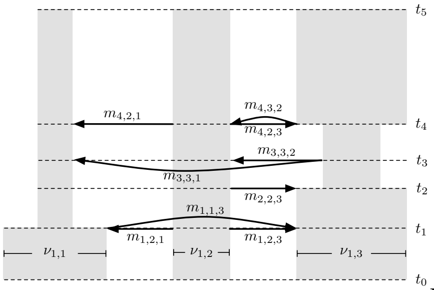


FIGURE 3. Accuracy and timing of Equation (4) in Box 2 for two populations of constant sizes $\nu_1 = 3$ and $\nu_2 = 2$ that diverged at time $t = 0.05$ from a population of size $\nu_A = 1$. No migration occurred after the split. Panels A, B, and C show a comparison between the approximate SFS (Equation 4) and the SFS generated by *momi*. Panels D-F show the same comparison, zoomed-in on the first 11×11 corner of the SFS corresponding to counts $i_1 = 0, ..., 10$ in population 1 and $i_2 = 0, ..., 10$ in population 2. Panels *G* through *I* show average errors over subsets of SFS terms. In Panels G-I, $\partial a \partial i$ and Equation (4) are separately compared to *momi*, which is taken as the true SFS. Averages are taken over subsets of SFS terms corresponding to the first 6×6 terms, the first 11×11 terms, and the first 21×21 terms. Panel *J* shows the percent error in each of the top 11×11 terms for a sample size of $n_1 = n_2 = n = 500$. Panel K shows the runtime for computing the first 11×11 corner of the SFS. Panel L shows the runtime for computing the full spectrum. Colors are consistent among panels A-F and J, and separately among panels G, H, I, K, and L. Plot marker shapes are consistent across panels.

Box 2 Approximating $\mathbb{E}\boldsymbol{\xi}_{\mathbf{n}}$ in P piecewise constant populations with pulse migrations.



Consider a collection of P piecewise constant populations with pulse migrations in which the relative size of population p is $\nu_p(t) = \nu_{k,p}$ for $t \in [t_{k-1}, t_k)$ measured in units of $2N$ generations and satisfying $t_0 < t_1 < \cdots < t_K < \infty$. Let $m_{k,p,p'}$ be the fraction of population p' that instantaneously immigrates from population p at the top of interval k, looking forward in time and let $\boldsymbol{M}_k$ be the $P \times P$ matrix whose (p, p') entry is $m_{k,p,p'}$ if $p \neq p'$ and $-\sum_{q=1,q\neq p}^{P} m_{k,q,p}$ if $p = p'$. Note that a population split at the top of interval k is handled by setting $m_{k,p,p'} = 1$ so that all lineages migrate into one population. Suppose that n_p sequences are sampled from population p, for $p = 1, \ldots, P$, and let $\mathbf{n} = (n_1, \ldots, n_P)$. Then entry $(i_1, \ldots, i_P)$ of the SFS $\boldsymbol{\xi}_{\mathbf{n}}$ can be computed using the formula

$$\mathbb{E}\xi_{\mathbf{n},(i_1,\ldots,i_P)} = \sum_{j_1,\ldots,j_P=0}^{i_1,\ldots,i_P} \mathbb{E}U_{\mathbf{n},(j_1,\ldots,j_P)} \prod_{k=1}^{P} (-1)^{i_k - j_k} \binom{i_k}{j_k} \binom{n_k}{i_k} \tag{4}$$

where

$$\mathbb{E}U_{\mathbf{n},(j_1,\ldots,j_P)} \approx \theta \sum_{k=1}^{K} \sum_{p=1}^{P} \nu_{k,p} \log \left[\frac{1 + [e^{\frac{t_k - t_{k-1}}{2\nu_{k,p}}} - 1]\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t_{k-1})}{1 + [e^{\frac{t_k - t_{k-1}}{2\nu_{k,p}}} - 1]\mathbb{E}\mathcal{A}_{\mathbf{n}-\mathbf{j},p}(t_{k-1})} \right]. \tag{5}$$

In Equation (4), the quantity $\mathbb{E}\mathcal{A}_{\mathbf{i},p}(t_k)$ is the expected number of ancestral lineages at time t_k in population p, before migration occurs looking forward in time, given that $\mathbf{i} = (i_1, \ldots, i_P)$ lineages are initially sampled at time zero. The quantity $\mathcal{A}_{\mathbf{i},p}(t_k)$ can be found recursively using

$$[\mathbb{E}\mathcal{A}_{\mathbf{i},1}(t_k), \ldots, \mathbb{E}\mathcal{A}_{\mathbf{i},P}(t_k)] = \boldsymbol{M}_k [\mathbb{E}\mathcal{A}'_{\mathbf{i},1}(t_k), \ldots, \mathbb{E}\mathcal{A}'_{\mathbf{i},P}(t_k)]^T \tag{6}$$

where

$$\mathbb{E}\mathcal{A}'_{\mathbf{i},p}(t_k) \approx \frac{\mathbb{E}\mathcal{A}_{\mathbf{i},p}(t_{k-1})}{\mathbb{E}\mathcal{A}_{\mathbf{i},p}(t_{k-1}) - [\mathbb{E}\mathcal{A}_{\mathbf{i},p}(t_{k-1}) - 1]e^{-(t_k - t_{k-1})/2\nu_{k,p}}} \tag{7}$$

is the number of lineages remaining in population p immediately after migration occurs at time t_k, looking forward in time, and $\mathbb{E}\mathcal{A}'_{\mathbf{i},p}(t_0) \equiv i_k$.

Considering all possible subsets $S_i^{(m)}$, $m = 1, \ldots, \binom{n}{i}$, of $\{1, \ldots, n\}$ with exactly i elements, we define a quantity $U_{n,i}$ as

$$U_{n,i} = \frac{1}{\binom{n}{i}} \sum_{m=1}^{\binom{n}{i}} \pi(S_i^{(m)}). \tag{8}$$

Intuitively, the quantity $U_{n,i}$ is the average number of sites that are private to a set of i lineages. Equation (8) is similar in form to the expression for the ith term of the SFS, which can be expressed as

$$\xi_{n,i} = \sum_{m=1}^{\binom{n}{2}} \alpha(S_i^{(m)}) \tag{9}$$

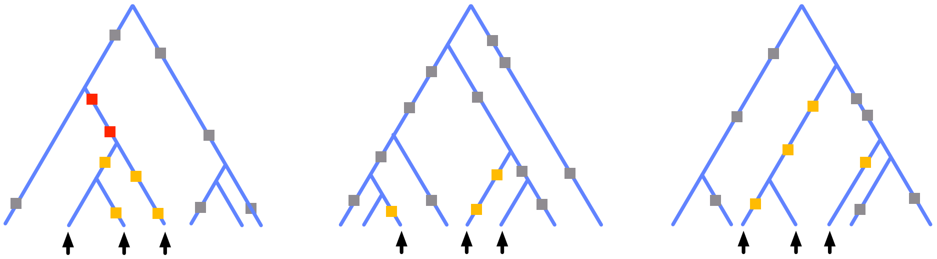


FIGURE 4. The quantities $\pi(S_i)$ and $\alpha(S_i)$. A subsample of $i = 3$ sequences (arrows) is shown for a total sample of $n = 7$ sequences. Trees are shown for three different loci along these sequences. Squares indicate mutations. For the indicated subsample, the number of private sites is $\pi(S_i) = 13$ (sum of all red and yellow squares), whereas the number of ancestral sites is $\alpha(S_i) = 2$ (red squares). Note that private sites are counted even when the subsample does not form a monophyletic clade.

By relating $\pi(S_i)$ with $\alpha(S_i)$, we can establish a relationship between $\mathbb{E}\xi_{n,i}$ with $\mathbb{E}U_{n,i}$. Because $\mathbb{E}U_{n,i}$ is proportional to an expected branch length, this approach allows us to establish a formula for $\mathbb{E}\boldsymbol{\xi}_n$ in terms of expected sums of branch lengths. This is the approach that we will take here to derive the approximations given Section 2.

3.1. **The relationship between U_n and $\boldsymbol{\xi}_n$ in one population.** We first establish the relationship between $U_{n,i}$ and $\xi_{n,i}$ in a single population before proceeding to the case of multiple populations. To derive the relationship, note that a site is private to S_i if and only if it is ancestral to some subset $S_j^{(q)}$ of $j \leq i$ sequences satisfying $S_j^{(q)} \leq S_i$, and to no others. Moreover, the set of mutations that are ancestral only to one particular subset $S_j^{(q)}$ of size j is mutually exclusive of the mutations ancestral only to a different subset $S_j^{(q')}$ of size j. Thus, the number $\pi(S_i)$ of private sites in S_i is equal to the sum over the ancestral sites in every subset of S_i of every size. In other words, we have

$$\pi(S_i) = \sum_{j=1}^{i} \sum_{q:S_j^{(q)} \subseteq S_i} \alpha(S_j^{(q)}). \tag{10}$$

Plugging Equation (10) into the definition in Equation (8) gives

$$\begin{aligned}
U_{n,i} &= \frac{1}{\binom{n}{i}} \sum_{m=1}^{\binom{n}{i}} \pi(S_i^{(m)}) \\
&= \frac{1}{\binom{n}{i}} \sum_{m=1}^{\binom{n}{i}} \sum_{j=1}^{i} \sum_{q=1}^{\binom{n}{j}} \alpha(S_j^{(q)}) \mathbf{I}[S_j^{(q)} \subseteq S_i^{(m)}] \\
&= \frac{1}{\binom{n}{i}} \sum_{j=1}^{i} \sum_{q=1}^{\binom{n}{j}} \alpha(S_j^{(q)}) \sum_{m=1}^{\binom{n}{i}} \mathbf{I}[S_j^{(q)} \subseteq S_i^{(m)}] \\
&= \frac{1}{\binom{n}{i}} \sum_{j=1}^{i} \sum_{q=1}^{\binom{n}{j}} \alpha(S_j^{(q)}) \binom{n-j}{i-j}
\end{aligned}$$

$$= \frac{1}{\binom{n}{i}} \sum_{j=1}^{i} \xi_{n,j} \binom{n-j}{i-j}$$

$$= \sum_{j=1}^{i} \frac{\binom{i}{j}}{\binom{n}{j}} \xi_{n,j}, \tag{11}$$

where $\mathbf{I}[\cdot]$ is the the indicator function, which is equal to 1 when its argument is true and 0 otherwise. The second equality in Equation (11) follows by plugging Equation (10) into Equation (8) and writing it in a slightly different way as a summation over all subsets of size i, times an indicator of set inclusion. The fourth equality follows from the fact that there are exactly $\binom{n-i}{j-i}$ sets of size i that contain a particular set of size j. This follows from the fact that, given that the j specific elements are in the set of size i, there are $\binom{n-j}{i-j}$ ways to choose the other $i-j$ elements of the set. The fifth equality follows from the definition of $\boldsymbol{\xi}_i$ in Equation (9) and the final equality follows from rearranging the identity $\binom{n}{j}\binom{n-j}{i-j} = \binom{i}{j}\binom{n}{i}$.

If we define $\boldsymbol{U}_n = (U_1, \ldots, U_{n-1})$ and $\boldsymbol{\xi}_n = (\xi_{n,1}, \ldots, \xi_{n,n-1})$, we can rewrite Equation (11) in matrix form to yield the particularly simple matrix relationship

$$\boldsymbol{U}_n = \mathbf{T}\boldsymbol{\xi}_n, \tag{12}$$

where $\mathbf{T}$ is the lower triangular matrix of dimension $(n-1)\times(n-1)$ whose element (i,j) is given by

$$[\mathbf{T}]_{ij} = \frac{\binom{i}{j}}{\binom{n}{j}}. \tag{13}$$

Because $\mathbf{T}$ is a lower-triangular matrix with non-zero diagonal elements ($[\mathbf{T}]_{ii} = \binom{n}{i}^{-1}$ on the diagonal), its determinant is nonzero and it is therefore invertible. As we will show in the more general case of multiple populations in Section 3.5, the inverse transformation from $\boldsymbol{U}$ to $\boldsymbol{\xi}$ in one population can be expressed as

$$\xi_{n,i} = \sum_{j=1}^{i} (-1)^{i-j} \binom{n}{i}\binom{i}{j} U_{n,j}, \tag{14}$$

which has the more compact matrix representation $\boldsymbol{\xi} = \mathbf{T}^{-1}\boldsymbol{U}$ where $\mathbf{T}^{-1}$ is the lower triangular matrix whose elements are given by

$$[\mathbf{T}^{-1}]_{ij} = (-1)^{i-j}\binom{n}{i}\binom{i}{j}. \tag{15}$$

3.2. Computing the expected number of private segregating sites in a single population.

We now obtain an expression for $\mathbb{E}U_{n,i} = \mathbb{E}\pi(S_i)$ in a single population by computing the expected number of sites that are private to a subset S_i. Suppose that n sequences are sampled at time $t = 0$ (the present) and let $\mathcal{A}_n(t)$ denote the random number of ancestral lineages remaining at time t in the past. Let $L_n(r,s)$ denote the sum of total branch lengths in the genealogy between times r and s in the past with $r \le s$. Then $L_n(r,s)$ can be expressed as an integral over the expected number of ancestral lineages:

$$\mathbb{E}L_n(r,s) = \int_r^s \mathbb{E}\mathcal{A}_n(t)dt. \tag{16}$$

The result in Equation (16) was stated as an asymptotic approximation by Chen and Chen (2013) in the limit as $n \to \infty$ (Equation 29 of that paper) and it was proved to hold exactly for finite n by Jewett and Rosenberg (2014) (Theorem 2.1 of that paper). Using Equation (16), Jewett and Rosenberg showed that under the infinite sites model, the expected total number of mutations $d_{[r,s]}(S_n)$ private to a set S_n of n homologous sequences arising during the time interval $[r, s]$ is given by

$$\mathbb{E}d_{[r,s]}(S_n) = \frac{\theta}{2}\mathbb{E}L_n(r,s) = \frac{\theta}{2}\int_r^s \mathbb{E}\mathcal{A}_n(t)dt. \tag{17}$$

Equation (17) can be used to compute the expected number of sites $\mathbb{E}\pi(S_i)$ that are private to a subset of i sampled sequences. In particular, if $d_{[r,s]}(S_n)$ is the total number of mutations private to the full sample of size n in the time interval $[r, s]$ and $d_{[r,s]}(S_n \backslash S_i)$ is the number of mutations private to the set $S_n \backslash S_i$ of $n - i$ other sequences, then

$$\pi_{[r,s]}(S_i) = d_{[r,s]}(S_n) - d_{[r,s]}(S_n \backslash S_i) \tag{18}$$

is the number of mutations arising in the time interval $[r, s]$ that are private to the sequences S_i, and no others. Taking the expectation of both sides and invoking Equation (17) gives

$$\mathbb{E}\pi_{[r,s]}(S_i) = \frac{\theta}{2}\mathbb{E}[L_n(r,s) - L_{n-i}(r,s)] = \frac{\theta}{2}\int_r^s [\mathbb{E}\mathcal{A}_n(t)dt - \mathbb{E}\mathcal{A}_{n-i}(t)]\, dt. \tag{19}$$

Equation (19) provides a simple way to compute expected numbers of private sites by integrating over expected numbers of ancestral lineages.

3.3. **Computing $\mathbb{E}\boldsymbol{\xi}_n$ exactly in a single population.** Equation (19) gives us a way to compute $U_{n,i}$, (and hence $\boldsymbol{\xi}_n$ using Equation 14) as long as we can compute the expected number of ancestors $\mathbb{E}\mathcal{A}_n(t)$ as a function of time. In a population with time-varying relative size $\nu(t)$, the expected number of ancestors can be computed exactly using the following expression due to Tavaré (1984) (Eqn. 5.11):

$$\mathbb{E}\mathcal{A}_n(t) = \sum_{m=1}^{n}(2m-1)\frac{n_{[m]}}{n_{(m)}}e^{-\binom{m}{2}\tau(t)}, \tag{20}$$

where $\tau(t)$ is the scaled coalescence time given by

$$\tau(t) = \int_0^t \frac{1}{\nu(z)}dz. \tag{21}$$

In Equation (20), the quantities $n_{[m]} = n(n-1)\cdots(n-m+1)$ and $n_{(m)} = n(n+1)\cdots(n+m-1)$ are the mth falling and rising factorials of n. If the population has constant size $\nu(r)$ in the time interval $[r, s]$, then integrating both sides of Equation (20) gives

$$\begin{aligned}\int_r^s \mathbb{E}\mathcal{A}_n(t)dt &= \sum_{m=1}^{n}(2m-1)\frac{n_{[m]}}{n_{(m)}}\int_r^s e^{-\binom{m}{2}\tau(t)}dt\\ &= (s-r) + \sum_{m=2}^{n}(2m-1)\frac{n_{[m]}}{n_{(m)}}\int_r^s e^{-\binom{m}{2}\tau(t)}dt\end{aligned}$$

$$
\begin{aligned}
&= (s-r) + \sum_{m=2}^{n} (2m-1) \frac{n_{[m]}}{n_{(m)}} \int_0^{s-r} e^{-\binom{m}{2}\left[\tau(r) + \frac{t}{\nu(r)}\right]} dt \\
&= (s-r) + \sum_{m=2}^{n} (2m-1) \frac{n_{[m]}}{n_{(m)}} \nu(r) e^{-\binom{m}{2}\tau(r)} \binom{m}{2}^{-1} \left[1 - e^{-\binom{m}{2}\frac{s-r}{\nu(r)}}\right]. \qquad (22)
\end{aligned}
$$

Combining Equation (22) with Equation (19) gives the expected number of private sites in a set of i sequences arising in the time interval $[r, s]$ in which the population has constant relative size $\nu(r)$:

$$
\begin{aligned}
\mathbb{E}\pi_{[r,s]}(S_i) &= \frac{\theta}{2} \int_r^s [\mathbb{E}\mathcal{A}_n(t) - \mathbb{E}\mathcal{A}_{n-i}(t)] dt \\
&= \frac{\theta}{2} \sum_{m=2}^{n} (2m-1) \frac{n_{[m]}}{n_{(m)}} \nu(r) e^{-\binom{m}{2}\tau(r)} \binom{m}{2}^{-1} \left[1 - e^{-\binom{m}{2}\frac{s-r}{\nu(r)}}\right] \\
&\quad - \frac{\theta}{2} \sum_{m=2}^{n-i} (2m-1) \frac{(n-i)_{[m]}}{(n-i)_{(m)}} \nu(r) e^{-\binom{m}{2}\tau(r)} \binom{m}{2}^{-1} \left[1 - e^{-\binom{m}{2}\frac{s-r}{\nu(r)}}\right] \\
&= \frac{\theta}{2} \sum_{m=2}^{n} (2m-1) \left[\frac{n_{[m]}}{n_{(m)}} - \frac{(n-i)_{[m]}}{(n-i)_{(m)}}\right] \nu(r) e^{-\binom{m}{2}\tau(r)} \binom{m}{2}^{-1} \left[1 - e^{-\binom{m}{2}\frac{s-r}{\nu(r)}}\right]. \qquad (23)
\end{aligned}
$$

In Equation (23), we were able to combine the two summations in the second equality because $(n-i)_{[m]}/(n-i)_{(m)} = 0$ for $m > n-i$; thus, the upper bound in the second summation can be set to n.

For a population history that is piecewise constant with relative size $\nu(t_{k-1})$ in each of the K epochs $\{[t_{k-1}, t_k)\}_{k=1}^{K}$, summing over all time intervals gives

$$
\mathbb{E}U_{n,i} = \frac{\theta}{2} \sum_{k=1}^{K} \sum_{m=2}^{n} (2m-1) \left[\frac{n_{[m]}}{n_{(m)}} - \frac{(n-i)_{[m]}}{(n-i)_{(m)}}\right] \nu(t_{k-1}) e^{-\binom{m}{2}\tau(t_{k-1})} \binom{m}{2}^{-1} \left[1 - e^{-\binom{m}{2}\frac{t_k - t_{k-1}}{\nu(t_{k-1})}}\right], \qquad (24)
$$

where $\tau(t_k) = \sum_{p'=1}^{k} \frac{t_{p'} - t_{p'-1}}{\nu(t_{p'})}$. Finally, taking the expectation of both sides of Equation (14) and plugging in Equation (24) gives

$$
\begin{aligned}
\mathbb{E}\xi_{n,i} &= \sum_{j=1}^{i} (-1)^{i-j} \binom{n}{i} \binom{i}{j} \mathbb{E}U_{n,j} \\
&= \frac{\theta}{2} \sum_{m=2}^{n} (2m-1) \sum_{j=1}^{i} (-1)^{i-j} \binom{n}{i} \binom{i}{j} \left[\frac{n_{[m]}}{n_{(m)}} - \frac{(n-j)_{[m]}}{(n-j)_{(m)}}\right] \\
&\quad \times \sum_{k=1}^{K} \nu(t_{k-1}) e^{-\binom{m}{2}\tau(t_{k-1})} \binom{m}{2}^{-1} \left[1 - e^{-\binom{m}{2}\frac{t_k - t_{k-1}}{\nu(t_{k-1})}}\right] \\
&= \frac{\theta}{2} \sum_{m=2}^{n} W_{n,i,m} \sum_{k=1}^{K} \nu(t_{k-1}) e^{-\binom{m}{2}\tau(t_{k-1})} \binom{m}{2}^{-1} \left[1 - e^{-\binom{m}{2}\frac{t_k - t_{k-1}}{\nu(t_{k-1})}}\right], \qquad (25)
\end{aligned}
$$

where

$$W_{n,i,m} = -(2m-1)\binom{n}{i}\sum_{j=0}^{i}(-1)^{i-j}\binom{i}{j}\frac{(n-j)_{[m]}}{(n-j)_{(m)}}. \tag{26}$$

In Equation (26), we have set the lower bound on the summation to $j = 0$ because $n_{[m]}/n_{(m)} - (n - j)_{[m]}/(n-j)_{(m)} = 0$ when $j = 0$. Moreover, the term $n_{[m]}/n_{(m)}$ in the second equality of Equation (25) drops out because it is constant in j and $\sum_{j=0}^{i}(-1)^{i-j}\binom{i}{j} = 0$ by the binomial theorem.

By comparison with Equation (10) of Kamm et al. (2017), it can be seen that the two formulas are the same. In particular, the quantity $W_{n,i,m}$ in Equation (26) is precisely the quantity $W_{n,i,m}$ in Kamm et al. (2017) and the internal summation over k in Equation (25) is the explicit form of the integral $\int_0^{t_K} e^{-\binom{m}{2}\frac{1}{\nu(t)}}dt$ for a population of constant size. Our approach provides a closed formula for the term $W_{n,i,m}$, which was found by a recursion in Kamm et al. (2017), following the derivation in Polanski and Kimmel (2003). However, because the recursion is faster to compute in practice, we provide the recursive form of $W_{n,i,m}$ in the final presentation of the formula in Box 1.

3.4. **Approximating $\mathbb{E}\boldsymbol{\xi}_n$ in a single population.** So far, we have computed exact formulas for the SFS. However, in preparation for deriving approximate formulas for the SFS that are computationally efficient in the case of multiple populations with pulse migration events, we first consider the approximation in a single population.

An approximate expression for $\xi_{n,i}$ can be obtained by following the the same approach used in Section 3.3, but replacing the exact formula for $\mathbb{E}\mathcal{A}_n(t)$ (Equation 20) with an approximate formula. The simplicity and computational efficiency of existing approximations of $\mathbb{E}\mathcal{A}_n(t)$ make it possible to obtain fast approximate formulas for $\mathbb{E}\pi_{[r,s]}(S_i)$, allowing us to obtain fast approximations of the SFS.

In a single population, Griffiths (1984) showed that the expected number of ancestors at time t in a population with relative size $\nu(t)$ for $t \in [0, \infty)$ can be approximated by

$$\mathbb{E}\mathcal{A}_n(t) \approx \frac{n}{n-(n-1)e^{-\tau(t)/2}}, \tag{27}$$

where $\tau(t)$ is the scaled coalescence time given in Equation (21). Griffiths showed that the approximation in Equation (27) holds asymptotically as $n \to \infty$ or as $t \to 0$.

If the population has constant size in the time interval $[r, s]$, then integrating both sides of Equation (27) over the time interval $[r, s]$ gives (Appendix C)

$$\int_r^s \mathbb{E}\mathcal{A}_n(t)dt \approx 2\nu(r)\log\left[1 + \frac{n[e^{\frac{s-r}{2\nu(r)}} - 1]}{n-(n-1)e^{-\tau(r)/2}}\right]. \tag{28}$$

If the population has piecewise constant size defined over the intervals $\{t_{k-1}, t_k\}_{k=1}^{K}$ in which $\nu(t) = \nu(t_{k-1})$ for $t \in [t_{k-1}, t_k)$ and $t_0 < t_1 < \cdots < t_K < \infty$, then by combining Equation (28) with Equation (19), we find that the number of sites private to i sequences that arise in the kth time interval is given by

$$\mathbb{E}\pi_{[t_{k-1},t_k]}(S_i) = \frac{\theta}{2}\int_{t_{k-1}}^{t_k}[\mathbb{E}\mathcal{A}_n(t) - \mathbb{E}\mathcal{A}_{n-i}(t)]dt$$

$$\approx \theta\nu(t_{k-1}) \log\left[\frac{Q(n, t_{k-1}, t_k)}{Q(n-i, t_{k-1}, t_k)}\right], \tag{29}$$

where $Q(n, r, s)$ is given by

$$Q(n, r, s) = 1 + \frac{n[e^{\frac{s-r}{2\nu(r)}} - 1]}{n - (n-1)e^{-\tau(r)/2}}, \tag{30}$$

and the scaled time in Equation (30) is given by

$$\tau(t) = \int_0^t \frac{1}{\nu(z)} dz = \sum_{k=1}^{K} \frac{t_k - t_{k-1}}{\nu(t_{k-1})}. \tag{31}$$

Summing Equation (29) over all time intervals and using the fact that $\mathbb{E}U_{n,i} = \mathbb{E}\pi(S_i)$ (Equation 8) gives

$$\mathbb{E}U_{n,i} \approx \theta \sum_{k=1}^{K} \nu(t_{k-1}) \log\left[\frac{Q(n, t_{k-1}, t_k)}{Q(n-i, t_{k-1}, t_k)}\right]. \tag{32}$$

Finally, plugging Equation (32) into the expectation of Equation (14) gives

$$\begin{aligned}
\mathbb{E}\xi_{n,i} &= \sum_{j=1}^{i} (-1)^{i-j} \binom{n}{i} \binom{i}{j} \mathbb{E}U_{n,j}, \\
&\approx \theta \sum_{k=1}^{K} \nu(t_{k-1}) \sum_{j=0}^{i} (-1)^{i-j} \binom{n}{i} \binom{i}{j} \log\left[\frac{Q(n, t_{k-1}, t_k)}{Q(n-j, t_{k-1}, t_k)}\right] \\
&= -\theta \sum_{k=1}^{K} \nu(t_{k-1}) \binom{n}{i} \sum_{j=0}^{i} (-1)^{i-j} \binom{i}{j} \log\left[Q(n-j, t_{k-1}, t_k)\right],
\end{aligned} \tag{33}$$

where the lower bound on the summation in the second equality can be taken to zero because the summand is zero at $j = 0$ and the final equality follows from the fact that $Q(n, r, s)$ is constant in j and $\sum_{j=0}^{i}(-1)^{i-j}\binom{i}{j} = 0$ by the binomial theorem. This gives the approximation of $\xi_{n,i}$ in Equation (3) of Box 1.

3.5. **The relationship between $\boldsymbol{U}_{\mathbf{n}}$ and $\boldsymbol{\xi}_{\mathbf{n}}$ in multiple populations.** The derivation of the relationship between $\boldsymbol{\xi}_{\mathbf{n}}$ and $\boldsymbol{U}_{\mathbf{n}}$ in multiple populations follows the same approach as the derivation in the case of a single population. In the case of P different populations with samples of sizes $n_1, \dots, n_P$, respectively, let $S_{i_1,\dots,i_P}$ denote a subsample of sequences with i_p lineages in population p, for $p = 1, \dots, P$. As in the case of a single population, define $\pi(S_{i_1,\dots,i_P})$ to be the number of private sites in the sample and define $\alpha(S_{i_1,\dots,i_P})$ to be the number of sites that are ancestral to the subsample and to no other sequences.

The multi-population forms of $U_{\mathbf{n},(i_1,\dots,i_P)}$ and $\xi_{\mathbf{n},(j_1,\dots,j_P)}$, are

$$U_{\mathbf{n},(i_1,\dots,i_P)} = \left[\prod_{p=1}^{P} \binom{n_p}{i_p}\right]^{-1} \sum_{m=1}^{\prod_{p=1}^{P}\binom{n_p}{i_p}} \pi(S^{(m)}_{i_1,\dots,i_P}), \tag{34}$$

and

$$\xi_{\mathbf{n},(i_1,\dots,i_P)} = \sum_{m=1}^{\prod_{p=1}^{P}\binom{n_p}{i_p}} \alpha(S^{(m)}_{i_1,\dots,i_P}). \tag{35}$$

The relationship between $U_{\mathbf{n},(i_1,\dots,i_P)}$ and $\xi_{\mathbf{n},(i_1,\dots,i_P)}$ can be established by the same approach we used to derive Equation (11). This derivation is carried out in Appendix A and gives

$$U_{\mathbf{n},(j_1,\dots,j_P)} = \sum_{i_1,\dots,i_P=0}^{j_1,\dots,j_P} \xi_{\mathbf{n},(i_1,\dots,i_P)} \prod_{p=1}^{P} \frac{\binom{j_p}{i_p}}{\binom{n_p}{i_p}}, \tag{36}$$

where we take $\boldsymbol{U}_{\mathbf{n},(0,\dots,0)} = \boldsymbol{\xi}_{\mathbf{n},(0,\dots,0)} = 0$. It is straightforward to check that the inverse transformation from $\boldsymbol{U}_{\mathbf{n}}$ to $\boldsymbol{\xi}_{\mathbf{n}}$ is given by

$$\xi_{\mathbf{n},(i_1,\dots,i_P)} = \sum_{j_1,\dots,j_P=0}^{i_1,\dots,i_P} U_{\mathbf{n},(j_1,\dots,j_P)} \prod_{p=1}^{P} (-1)^{i_p-j_p} \binom{n_p}{i_p}\binom{i_p}{j_p}, \tag{37}$$

which can be checked by plugging Equation (37) into Equation (36) and showing that the composite transformation yields the identity. We carry out these calculations in Appendix B.

3.6. Approximating $\boldsymbol{U}_{\mathbf{n}}$ in multiple populations with piecewise constant sizes and pulse migrations.

The results of Sections 3.2 through 3.5 can be combined to obtain a fast approximate formula for $\boldsymbol{\xi}_{\mathbf{n}}$ in a collection of piecewise constant populations connected by pulse migrations. In particular, for a set of P populations with n_p lineages sampled from population p, for $p = 1,\dots,P$, let $\mathcal{A}_{\mathbf{n}}(t)$ denote the total number of ancestors at time t of the set $S_{n_1,\dots,n_P}$ of $n_1,\dots,n_P$ sampled sequences. Similarly, let $\mathcal{A}_{\mathbf{i}}(t)$ denote the total number of ancestors at time t of a subset $S_{i_1,\dots,i_P}$ of $i_1,\dots,i_P$ sequences. Finally, let $\mathcal{A}_{\mathbf{n},p}(t)$ denote the number of ancestors of $n_1,\dots,n_P$ sequences that exist in population p at time t, and similarly define $\mathcal{A}_{\mathbf{i},p}(t)$.

If the size of each of the P populations is constant in a time interval $[r,s]$, then the total number of alleles arising in the time interval $[r,s]$ that are private to the subsample $S_{i_1,\dots,i_P}$ can be found using Equation (19) as

$$\begin{aligned}
&\mathbb{E}\pi_{[r,s]}(S_{i_1,\dots,i_P})\\
&= \frac{\theta}{2}\int_r^s [\mathbb{E}\mathcal{A}_{\mathbf{n}}(t) - \mathbb{E}\mathcal{A}_{\mathbf{n}-\mathbf{i}}(t)]\,dt\\
&= \frac{\theta}{2}\sum_{p=1}^{P}\int_r^s [\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t) - \mathbb{E}\mathcal{A}_{\mathbf{n}-\mathbf{i},p}(t)]\,dt\\
&\approx \theta\sum_{p=1}^{P}\nu_p(r)\log\left[\frac{1+[e^{\frac{s-r}{2\nu_p(r)}}-1]\mathbb{E}\mathcal{A}_{\mathbf{n},p}(r)}{1+[e^{\frac{s-r}{2\nu_p(r)}}-1]\mathbb{E}\mathcal{A}_{\mathbf{n}-\mathbf{i},p}(r)}\right],
\end{aligned} \tag{38}$$

where the final equality comes from Equation (C.5) in the Appendix.

Now suppose that the relative size of the pth population is $\nu_p(t) = \nu_p(t_{k-1})$ in the time interval $t \in [t_{k-1}, t_k)$, for K different time intervals $\{[t_{k-1},t_k)\}_{k=1}^{K}$ satisfying $t_0 < \cdots < t_K < \infty$. Then the total number of private segregating sites over all time intervals is

$$\mathbb{E}U_{\mathbf{n},(i_1,\ldots,i_P)} \equiv \mathbb{E}\pi(S_{i_1,\ldots,i_P}) \approx \theta \sum_{k=1}^{K}\sum_{p=1}^{P} \nu_p(t_{k-1}) \log\left[\frac{1+[e^{\frac{t_k-t_{k-1}}{2\nu_p(t_{k-1})}}-1]\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t_{k-1})}{1+[e^{\frac{t_k-t_{k-1}}{2\nu_p(t_{k-1})}}-1]\mathbb{E}\mathcal{A}_{\mathbf{n}-\mathbf{i},p}(t_{k-1})}\right]. \tag{39}$$

To find $\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t)$, we can use the fact that pulse migrations simply transfer ancestral lineages from one population to another at fixed times in the past. Let $m_{k,p,p'}$ be the fraction of population p' that instantaneously immigrates from population p at the top of interval k, looking forward in time and let $\boldsymbol{M}_k$ be the $P \times P$ matrix whose (p,p') entry is $[\boldsymbol{M}_k]_{p,p'} = m_{k,p,p'}$ if $p \neq p'$ and $[\boldsymbol{M}_k]_{p,p} = -\sum_{q=1,q\neq p}^{P} m_{k,q,p}$. Given that $\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t_{k-1})$ ancestors exist in population p at the bottom of the kth time interval, then using Equation (27), the number $\mathbb{E}\mathcal{A}'_{\mathbf{n},p}(t_k)$ remaining at the top of the interval immediately before migration is approximately

$$\mathbb{E}\mathcal{A}'_{\mathbf{n},p}(t_k) \approx \frac{\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t_{k-1})}{\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t_{k-1}) - [\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t_{k-1})-1]e^{-\frac{t_k-t_{k-1}}{2\nu_p(t_{k-1})}}} \tag{40}$$

and the number remaining at the top after migration is

$$\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t_k) = \sum_{p'=1}^{P} m_{k,p,p'}\mathbb{E}\mathcal{A}'_{\mathbf{n},p'}(t_k), \tag{41}$$

which has the more compact matrix representation

$$(\mathbb{E}\mathcal{A}_{\mathbf{n},1}(t_k),\ldots,\mathbb{E}\mathcal{A}_{\mathbf{n},P}(t_k)) = \boldsymbol{M}_k(\mathbb{E}\mathcal{A}'_{\mathbf{n},1}(t_k),\ldots,\mathbb{E}\mathcal{A}'_{\mathbf{n},P}(t_k)). \tag{42}$$

Thus, the quantities $\mathbb{E}\mathcal{A}_{\mathbf{n},p}(t_k)$ can be found by recursively applying Equations (40) and (42). Applying the relationship between $\xi_{\mathbf{n},(j_1,\ldots,j_P)}$ and $U_{\mathbf{n},(i_1,\ldots,i_P)}$ given in Equation (37) to Equation (39) gives the result in Equation (4) of Box 2.

4. Discussion

We have obtained accurate approximate formulas for computing the SFS in populations of piecewise constant size with instantaneous pulse migrations among them. The computational complexity of these formulas depends the index of the term of the SFS being computed, rather than on the sample size, allowing low-order terms of the SFS to be computed quickly for arbitrarily large sample sizes.

Our formulas for the SFS were derived by conceptualizing the SFS as a weighted sum over expected total ancestral branch lengths among i sampled lineages. In contrast, previous approaches expressed the SFS as a weighted sum over expected first coalescence times among $2,\ldots,n$ lineages (Polanski and Kimmel 2003, Chen and Chen 2013, Kamm et al. 2017). Conceptualizing the SFS in terms of expected sums of branch lengths makes it possible to obtain the simple and fast formulas we derive here. The approach is useful more generally for deriving approximate coalescent formulas under complicated demographic models (Jewett and Rosenberg 2014).

It is important to note that approaches based on the diffusion approximation of the Wright-Fisher model are still the most efficient methods for computing the full SFS when the number of populations is small. However, computing the full SFS becomes intractably slow as the number of samples and populations increases. Thus, the benefit of the approximate formulas we present is their higher efficiency for computing a subset of SFS terms when sample sizes are large. The approximations derived here are also more accurate than the diffusion approximation for low order terms of the SFS when the number of sampled haplotypes is moderate or large.

The formulas we obtain are for populations of piecewise constant size with instantaneous pulse migrations. However, they can be used to approximate the SFS for populations of time-varying size and continuous migration by taking the time-step to be short. It is also possible to derive approximations of the SFS in the case of exponentially growing populations and continuous migration by substituting approximate or exact formulas for the expected number of ancestral lineages under these scenarios into the penultimate equality of Equation (38). Approximations for the expected number of ancestral lineages under continuous migration and arbitrary size changes are given in Jewett and Rosenberg (2014). However, for populations with exponentially changing sizes the approach described here yields formulas that are computationally less efficient and numerically less stable than existing methods. Thus, we have chosen to focus on the fast approximations for piecewise constant populations presented in this paper.

5. Acknowledgments

This research was supported in part by a National Institutes of Health grants R01-GM109454 and R01-GM094402 and by a Packard Fellowship for Science and Engineering awarded to Yun S. Song. I am very grateful to Yun Song for his support and multiple close readings of the manuscript. Many thanks also to Jeffrey Spence for lots of helpful discussions and suggestions.

References

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

H. Chen. The joint allele frequency spectrum of multiple populations: a coalescent theory approach. *Theor. Popul. Biol.*, 81(2):179–195, 2012.

H. Chen and K. Chen. Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. *Genetics*, 194:721–736, 2013.

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

J.C. Fay and C.I. Wu. Hitchhiking under positive darwinian selection. *Genetics*, 155(3):1405–1413, 2000.

S. Gravel, B.M. Henn, R.N. Gutenkunst, A.R. Indap, G.T. Marth, A.G. Clark, F. Yu, R.A. Gibbs, 100Genomes, and C.D. Bustamante. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci., USA*, 108(29):11983–11988, 2011.

R.C. Griffiths. Asymptotic line-of-descent distributions. *J. Math. Biology*, 21:67–75, 1984.

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

D. L. Hartl and A. G. Clark. *Principles of Population Genetics, 4th ed.* Sinauer Associates, 2007.

E.M. Jewett and N.A. Rosenberg. Theory and applications of a deterministic approximation to the coalescent model. *Theor. Popul. Biol.*, 93:14–29, 2014.

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

J. Kamm, J. Terhorst, and Y.S. Song. Efficient computation of the joint sample frequency spectra for multiple populations. *J. Comput. Graph. Stat.*, 26:182–194, 2017.

A. Keinan, J.C. Mullikin, N. Patterson, and D. Reich. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.*, 39 (10):1251–1255, 2007.

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

G.T. Marth, E. Czabarka, J. Murvai, and S.T. Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–372, 2004.

Y.E. Maruvka, N.M. Shnerb, Y. Bar-Yam, and J. Wakeley. Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate. *Mol. Biol. Evol.*, 28:1617–1631, 2011.

R.C. McCoy, N.R. Garud, J.L. Kelley, C.L. Boggs, and D.A. Petrov. Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population. *Mol. Ecol.*, 23(1):136–150, 2014.

R. Nielsen, C.D. Bustamante, A.G. Clark, S. Glanowski, T.B. Sackton, M.J. Hubisz, A. Fledel-Alon, D.M. Tanenbaum, D. Civello, T.J. White, J.J. Sninsky, M.D. Adams, and M. Cargill. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*, 3(6):e170, 2005.

A. Polanski and M. Kimmel. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1): 427–436, 2003.

M. Slatkin and B. Rannala. Estimating the age of alleles by use of intraallelic variability. *Am. J. Hum. Genet.*, 60:447–458, 1997.

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

S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, 26:119–164, 1984.

E.M. Volz, S.L. Kosakovsky Pond, M.J. Ward, A.J. Leigh Brown, and S.D.W. Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183:1421–1430, 2009.

J. Wakeley. *Coalescent theory: An introduction.* Roberts & Company Publishers, Greenwood Village, CO, 2008.

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

G.A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7:256–276, 1975.

Appendix A. The relationship between $U_{\mathbf{n},(j_1,\dots,j_P)}$ and $\xi_{\mathbf{n},(i_1,\dots,i_P)}$

To derive the relationship between $U_{\mathbf{n},(j_1,\dots,j_P)}$ and $\xi_{\mathbf{n},(i_1,\dots,i_P)}$, we begin with the definition of $U_{\mathbf{n},(j_1,\dots,j_P)}$:

$$\begin{aligned}
U_{\mathbf{n},(j_1,\dots,j_P)} &= \left[\frac{1}{\prod_{p=1}^P \binom{n_p}{j_p}}\right] \sum_{m=1}^{\prod_{p=1}^P \binom{n_p}{j_p}} \pi(S^{(m)}_{j_1,\dots,j_P}) \\
&= \left[\frac{1}{\prod_{p=1}^P \binom{n_p}{j_p}}\right] \sum_{m=1}^{\prod_{p=1}^P \binom{n_p}{j_p}} \sum_{i_1,\dots,i_P=0}^{j_1,\dots,j_P} \sum_{h=1}^{\prod_{p=1}^P \binom{n_p}{i_p}} \alpha(S^{(h)}_{i_1,\dots,i_P})\mathbf{I}[S^{(h)}_{i_1,\dots,i_P} \subseteq S^{(m)}_{j_1,\dots,j_P}] \\
&= \left[\frac{1}{\prod_{p=1}^P \binom{n_p}{j_p}}\right] \sum_{i_1,\dots,i_P=0}^{j_1,\dots,j_P} \sum_{h=1}^{\prod_{p=1}^P \binom{n_p}{i_p}} \alpha(S^{(h)}_{i_1,\dots,i_P}) \sum_{m=1}^{\prod_{p=1}^P \binom{n_p}{j_p}} \mathbf{I}[S^{(h)}_{i_1,\dots,i_P} \subseteq S^{(m)}_{j_1,\dots,j_P}] \\
&= \left[\frac{1}{\prod_{p=1}^P \binom{n_p}{j_p}}\right] \sum_{i_1,\dots,i_P=0}^{j_1,\dots,j_P} \sum_{h=1}^{\prod_{p=1}^P \binom{n_p}{i_p}} \alpha(S^{(h)}_{i_1,\dots,i_P}) \prod_{p=1}^P \binom{n_p - i_p}{j_p - i_p} \\
&= \left[\frac{1}{\prod_{p=1}^P \binom{n_p}{j_p}}\right] \sum_{i_1,\dots,i_P=0}^{j_1,\dots,j_P} \xi_{\mathbf{n},(i_1,\dots,i_P)} \prod_{p=1}^P \binom{n_p - i_p}{j_p - i_p} \\
&= \sum_{i_1,\dots,i_P=0}^{j_1,\dots,j_P} \xi_{\mathbf{n},(i_1,\dots,i_P)} \prod_{p=1}^P \frac{\binom{j_p}{i_p}}{\binom{n_p}{i_p}}. \qquad \text{(A.1)}
\end{aligned}$$

In Equation (A.1), the second equality follows from writing the summand as a sum over alleles ancestral to all subsets $S^{(h)}_{i_1,\dots,i_P} \subseteq S^{(m)}_{j_1,\dots,j_P}$ such that $i_p \leq j_p$ for $p = 1,\dots,P$. The fourth equality follows from the fact that $\binom{n_p-i_p}{j_p-i_p}$ is the number of subsets of size j_p in population p that contain a particular subset of size i_p. As in the single-population case, this result follows from the fact that there are $\binom{n_p-i_p}{j_p-i_p}$ ways to choose the $j_p - i_p$ other members of this subset. The fifth equality follows from the definition of $\xi_{\mathbf{n},(i_1,\dots,i_P)}$ in Equation (35) and the final equality follows from rearranging the identity $\binom{n}{i}\binom{n-i}{j-i} = \binom{j}{i}\binom{n}{j}$.

Appendix B. The inverse transform from $U_{\mathbf{n},(j_1,\dots,j_P)}$ to $\xi_{\mathbf{n},(i_1,\dots,i_P)}$

This transform can be derived by plugging Equation (37) into Equation (36) and showing that the composite transformation yields the identity:

$$\begin{aligned}
U_{\mathbf{n},(j_1,\dots,j_P)} &= \sum_{i_1,\dots,i_P=0}^{j_1,\dots,j_P} \xi_{\mathbf{n},(i_1,\dots,i_P)} \prod_{p=1}^P \frac{\binom{j_p}{i_p}}{\binom{n_p}{i_p}} \\
&= \sum_{i_1,\dots,i_P=0}^{j_1,\dots,j_P} \sum_{m_1,\dots,m_P=0}^{i_1,\dots,i_P} U_{\mathbf{n},(m_1,\dots,m_P)} \prod_{p=1}^P (-1)^{i_p-m_p} \binom{n_p}{i_p}\binom{i_p}{m_p}\binom{j_p}{i_p}\binom{n_p}{i_p}^{-1} \\
&= \sum_{i_1,\dots,i_P=0}^{j_1,\dots,j_P} \sum_{m_1,\dots,m_P=0}^{i_1,\dots,i_P} U_{\mathbf{n},(m_1,\dots,m_P)} \prod_{p=1}^P (-1)^{i_p-m_p} \binom{j_p}{i_p}\binom{i_p}{m_p}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i_1,\ldots,i_P=0}^{j_1,\ldots,j_P} \sum_{m_1,\ldots,m_P=0}^{j_1,\ldots,j_P} U_{\mathbf{n},(m_1,\ldots,m_P)} \prod_{p=1}^{P} (-1)^{i_p-m_p} \binom{j_p}{i_p}\binom{i_p}{m_p} \\
&= \sum_{m_1,\ldots,m_P=0}^{j_1,\ldots,j_P} U_{\mathbf{n},(m_1,\ldots,m_P)} \prod_{p=1}^{P} \sum_{i_p=m_p}^{j_p} (-1)^{i_p-m_p} \binom{j_p}{m_p}\binom{j_p-m_p}{i_p-m_p} \\
&= \sum_{m_1,\ldots,m_P=0}^{j_1,\ldots,j_P} U_{\mathbf{n},(m_1,\ldots,m_P)} \prod_{p=1}^{P} \binom{j_p}{m_p} \sum_{i_p=m_p}^{j_p} (-1)^{i_p-m_p} \binom{j_p-m_p}{i_p-m_p} \\
&= \sum_{m_1,\ldots,m_P=0}^{j_1,\ldots,j_P} U_{\mathbf{n},(m_1,\ldots,m_P)} \prod_{p=1}^{P} \mathbf{I}[m_p = j_p]. \qquad \text{(B.1)}
\end{aligned}
$$

In the fourth equality, we have extended the upper bound of the inner summation up to $j_1, \ldots, j_P$ using the fact that $\binom{i_p}{m_p} = 0$ for $m_p > i_p$. In the fifth equality, we brought the summation inside the product and used the identity $\binom{p}{r}\binom{r}{q} = \binom{p}{q}\binom{p-q}{r-q}$. The final equality follows from reindexing $h_p = i_p - m_p$ and noting that the summation $\sum_{h_p=0}^{j_p-m_p}(-1)^{h_p}\binom{j_p-m_p}{h_p} = (1-1)^{j_p-m_p} = 0$ whenever $j_p \neq m_p$ by the binomial theorem, and it is equal to one if $j_p = m_p$. Thus, we arrive at the fact that the right-hand side of Equation (B.1) is equal to $U_{\mathbf{n},(j_1,\ldots,j_P)}$, proving the identity.

Appendix C. Approximation of $\int_r^s \mathbb{E}\mathcal{A}_n(t)dt$

The derivation of Equation (28) amounts to a change of variables and some algebra. First, noting that $\tau(t) = \tau(r) + (t-r)/\nu(r)$ for $t \in [r, s]$ whenever the relative population size is constant in $[r, s]$, we have

$$
\begin{aligned}
&\int_r^s \mathbb{E}\mathcal{A}_n(t)dt \\
&\approx \int_r^s \frac{n}{n-(n-1)e^{-\tau(t)/2}} dt \\
&\approx \int_r^s \frac{n}{n-(n-1)e^{-\tau(r)/2+r/2\nu(r)-t/2\nu(r)}} dt. \qquad \text{(C.1)}
\end{aligned}
$$

Thus, setting $b = -(n-1)e^{-\tau(r)/2+r/2\nu(r)}$, $c = -1/2\nu(r)$ we have an integral of the form

$$
n \int_r^s \frac{1}{n+be^{ct}} dt. \qquad \text{(C.2)}
$$

Making the change of variables $y = e^{ct}$ so that $dt = \frac{1}{cy}dy$, we have

$$
\begin{aligned}
&n \int_r^s \frac{1}{n+be^{ct}} dt \\
&= n \int_{y=e^{cr}}^{e^{cs}} \frac{1}{n+by}\frac{1}{cy} dy \\
&= \frac{1}{c} \log\left[\frac{y}{n+by}\right]\Bigg|_{e^{cr}}^{e^{cs}}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{c}\log\left[\frac{1}{ny^{-1}+b}\right]\Bigg|_{e^{cr}}^{e^{cs}} \\
&= -\frac{1}{c}\log\left[\frac{ne^{-cs}+b}{ne^{-cr}+b}\right] \\
&= -\frac{1}{c}\log\left[\frac{ne^{-c(s-r)}+be^{cr}}{n+be^{cr}}\right] \\
&= -\frac{1}{c}\log\left[\frac{ne^{-c(s-r)}-n+n+be^{cr}}{n+be^{cr}}\right] \\
&= -\frac{1}{c}\log\left[1+\frac{n[e^{-c(s-r)}-1]}{n+be^{cr}}\right]. \qquad \text{(C.3)}
\end{aligned}
$$

Plugging in $b = -(n-1)e^{-\tau(r)/2+r/2\nu(r)}$ and $c = -1/2\nu(r)$, we obtain the result in Equation (28):

$$
\int_r^s \mathbb{E}\mathcal{A}_n(t)dt \approx 2\nu(r)\log\left[1+\frac{n[e^{\frac{s-r}{2\nu(r)}}-1]}{n-(n-1)e^{-\tau(r)/2}}\right], \qquad \text{(C.4)}
$$

Noting that $n/(n-(n-1)e^{-\tau(r)/2}) \equiv \mathcal{A}_n(r)$, we can express Equation (C.4) more compactly as

$$
\int_r^s \mathbb{E}\mathcal{A}_n(t)dt \approx 2\nu(r)\log\left[1+[e^{\frac{s-r}{2\nu(r)}}-1]\mathbb{E}\mathcal{A}_n(r)\right]. \qquad \text{(C.5)}
$$

Equation (C.5) allows us to approximate the expected branch length in a time interval $[r,s]$ as long as we know the number of ancestors remaining at time r.