The Site Frequency Spectrum for General Coalescents

Jeffrey P. Spence,*,[1] John A. Kamm,[†,1] and Yun S. Song[†,‡,§,,††,2]**

*Computational Biology Graduate Group, [†]Department of Statistics, [‡]Computer Science Division, and [§]Department of Integrative Biology, University of California, Berkeley, California 94720, and **Department of Mathematics and [††]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

ABSTRACT General genealogical processes such as Λ- and Ξ-coalescents, which respectively model multiple and simultaneous mergers, have important applications in studying marine species, strong positive selection, recurrent selective sweeps, strong bottlenecks, large sample sizes, and so on. Recently, there has been significant progress in developing useful inference tools for such general models. In particular, inference methods based on the site frequency spectrum (SFS) have received noticeable attention. Here, we derive a new formula for the expected SFS for general Λ- and Ξ-coalescents, which leads to an efficient algorithm. For time-homogeneous coalescents, the runtime of our algorithm for computing the expected SFS is $O(n^2)$, where n is the sample size. This is a factor of n^2 faster than the state-of-the-art method. Furthermore, in contrast to existing methods, our method generalizes to time-inhomogeneous Λ- and Ξ-coalescents with measures that factorize as $\Lambda(dx)/\zeta(t)$ and $\Xi(d\mathbf{x})/\zeta(t)$, respectively, where ζ denotes a strictly positive function of time. The runtime of our algorithm in this setting is $O(n^3)$. We also obtain general theoretical results for the identifiability of the Λ measure when ζ is a constant function, as well as for the identifiability of the function ζ under a fixed Ξ measure.

KEYWORDS frequency spectrum; identifiability; multiple merger; simultaneous merger

WHEN summarizing sequence data from n individuals, a natural and often-used statistic is the site frequency spectrum (SFS), $\hat{\tau}_n = (\hat{\tau}_{n,1}, \ldots, \hat{\tau}_{n,n-1})^T$, where $\hat{\tau}_{n,k}$ is simply the number of sites at which k of n individuals carry the mutant (or the derived) allele. Despite being only $n-1$ numbers, the SFS still contains a surprising amount of information about the history and structure of the population from which the individuals were sampled. Indeed, for neutrally evolving populations that are well modeled by Kingman's coalescent (Kingman 1982), the expected value of the SFS was first computed for populations of constant size (Fu 1995), extended to populations of variable size (Griffiths and Tavaré 1998; Polanski *et al.* 2003; Polanski and Kimmel 2003), and has since been used as a statistic for demographic inference in numerous studies (*e.g.*, Nielsen 2000; Gutenkunst *et al.* 2009; Coventry *et al.* 2010; Gravel *et al.* 2011; Excoffier *et al.* 2013; Bhaskar *et al.* 2015; Kamm *et al.* 2015; Gao and Keinan 2016).

Yet not all populations are well modeled by Kingman's coalescent. In fact, Kingman's coalescent can be viewed as a special case of a broader class of coalescent processes called Λ-coalescents (Pitman 1999; Sagitov 1999). While Kingman's coalescent permits only pairwise mergers of lineages, Λ-coalescents allow two or more lineages to merge simultaneously in a single coalescence event. Such events arise when a single individual has many offspring (Möhle and Sagitov 2001; Eldon and Wakeley 2006), under models of recurrent selective sweeps (Durrett and Schweinsberg 2004, 2005), in populations undergoing continuous strong selection (Neher and Hallatschek 2013; Schweinsberg 2015), and in many other models. Λ-Coalescents can further be seen as special cases of a broader class of coalescents called Ξ-coalescents (Schweinsberg 2000). In Ξ-coalescents, more than one merger event can occur simultaneously, resulting in simultaneous multiple mergers. While Ξ-coalescents have received less attention than Λ-coalescents in the literature, they still arise in certain models of selection (Huillet 2014), models of selective sweeps (Durrett and Schweinsberg 2005), models with repeated strong bottlenecks (Birkner *et al.* 2009), and for certain

doi: 10.1534/genetics.115.184101
Manuscript received October 26, 2015; accepted for publication February 10, 2016; published Early Online February 12, 2016.
[1]These authors contributed equally to this work.
[2]Corresponding author: Department of Statistics, 321 Evans Hall #3860, University of California, Berkeley, Berkeley, CA 94720-3860. E-mail: yss@berkeley.edu

diploid mating models (Möhle and Sagitov 2003). Also, since Ξ-coalecents generalize Λ-coalescents, any results presented about Ξ-coalescents immediately pertain to Λ-coalescents.

More formally, time-homogeneous Ξ-coalescents are governed by a measure $\Xi(d\mathbf{x})$ on the set $\{(x_1, x_2, \ldots) : x_1 \geq x_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} x_i \leq 1\}$ Furthermore, we consider time-inhomogeneous Ξ-coalescents with measures that decompose into a time-independent part $\Xi(d\mathbf{x})$ and a strictly positive function $\zeta : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$ of time, where $\zeta(t)$ represents (for historical reasons) the inverse intensity. That is, the coalescent is now governed by the measure $\Xi(d\mathbf{x})/\zeta(t)$. For example, for Kingman's coalescent, $\Xi(d\mathbf{x}) = \delta_{\mathbf{0}}(d\mathbf{x})$, the point mass at zero, and $\zeta(t)$ corresponds to the scaled effective population size at time t. For other models, $\zeta(t)$ does not necessarily correspond to the population size, but has an interpretation specific to the model. For example, Neher and Hallatschek (2013) show empirically that the rate of coalescence events in a model of continuous strong selection is a nonlinear function of the population size and the first two moments of the distribution of mutational effects. For a review of the mechanics of Λ-coalescents, see Pitman (1999) and for a review of Ξ-coalescents, see Schweinsberg (2000). For an alternative perspective, see Donnelly and Kurtz (1999) and Birkner *et al.* (2009) for a lookdown construction of particle systems with general reproduction mechanisms.

As mentioned above, the expected SFS for Kingman's coalescent is well understood and can, in fact, be computed for an arbitrary ζ in $O(n^2)$ time (Polanski and Kimmel 2003). For Λ- and Ξ-coalescents, however, the expected SFS can be computed only for constant ζ and the method for Λ-coalescents takes $O(n^4)$ time (Birkner *et al.* 2013a) and the method for Ξ-coalescents takes time exponential in n as a sum over partitions of the first n numbers must be performed (Blath *et al.* 2015). Here we present a method that can compute the expected SFS for time-inhomogeneous Λ- and Ξ-coalescents with arbitrary ζ in $O(n^3)$ time. In the case where ζ is a constant function, our method can compute the expected SFS in $O(n^2)$ time given the rate matrix $\mathbf{Q}$ of the ancestral process, which is defined more precisely below. We also prove some results about the sample size needed to make Λ identifiable for popular classes of Λ measures for constant ζ, as well as results about the sample size needed to make ζ identifiable for a fixed $\Xi(d\mathbf{x})$.

There has also been some related work on determining the asymptotic behavior of the expected SFS as $n \to \infty$. In this setting, Berestycki *et al.* (2007, 2014) derive some simple formulas for time-homogeneous Λ-coalescents that come down from infinity. For finite n, however, these asymptotic formulas can be rather inaccurate. Indeed, even for $n = 10,000$, Birkner *et al.* (2013a) show that for some Λ-coalescents, there is a sizable discrepancy between the asymptotic formulas and the SFS obtained by simulation, highlighting the need for finite-sample calculations. Nevertheless, such asymptotic results highlight some interesting properties of Λ-coalescents and are reviewed in Berestycki (2009).

The remainder of this article is organized as follows. We first present our main results about the computation of the SFS for time-inhomogeneous coalescents and discuss the practical runtime of our implementation. We also investigate the variation in the empirical SFS and study the ability to infer the underlying model, using the empirical SFS. Then, we prove some identifiability results about general coalescents. We conclude with a discussion on the implications of our results.

Main Theoretical Results on the Expected SFS

Here we present our theoretical results on the expected SFS for a general Ξ-coalescent with a measure of the form $\Xi(d\mathbf{x})/\zeta(t)$. These results lead to an $O(n^3)$-time algorithm for computing the expected SFS and can be improved to $O(n^2)$ if ζ is a constant function. Briefly, we use subsampling arguments to show that the expected SFS $\boldsymbol{\tau}_n = \mathbb{E}[\hat{\boldsymbol{\tau}}_n]$ can be computed from $\mathbf{a}_n := (\mathbb{E}T_2^{\mathrm{MRCA}}, \ldots, \mathbb{E}T_n^{\mathrm{MRCA}})^T$, where $\mathbb{E}T_k^{\mathrm{MRCA}}$ denotes the expected time to the most recent common ancestor for sample size $k \in \{2, \ldots, n\}$. Then, we show how to compute $\mathbf{a}_n$, using a spectral decomposition of the rate matrix $\mathbf{Q}$ of the ancestral process (also known as the block-counting process) of the time-homogeneous coalescent corresponding to $\Xi(d\mathbf{x})$. More specifically, $\mathbf{Q}$ is a lower triangular matrix, where $(\mathbf{Q})_{ij}$ is the instantaneous rate at which i unlabeled lineages merge to form j unlabeled lineages when $\zeta \equiv 1$. For example, for Kingman's coalescent,

$$(\mathbf{Q})_{ij} = \begin{cases} \binom{i}{2}, & j = i-1, \\ -\binom{i}{2}, & j = i, \\ 0, & \text{otherwise.} \end{cases}$$

Using this notation, we are now ready to state our main result. The rest of this section provides lemmas that contain formulas for the matrices in *Theorem 1*, as well as a proof of those lemmas and *Theorem 1*.

Theorem 1. *Consider an arbitrary time-inhomogeneous Ξ-coalescent governed by a measure $\Xi(d\mathbf{x})/\zeta(t)$, such that the expected time $c_{k,k}$ to the first coalescence for a sample of size k is finite for $k \in \{2, \ldots, n\}$. Let $\mathbf{c}_n = (c_{2,2}, \ldots, c_{n,n})^T$. Then, there exists a universal matrix $\mathbf{A} \in \mathbb{R}^{n-1 \times n-1}$ that does not depend on the measure and a matrix $\mathbf{L} \in \mathbb{R}^{n-1 \times n-1}$ that depends on Ξ but not ζ, such that*

$$\boldsymbol{\tau}_n = \frac{\theta}{2}\mathbf{A}\mathbf{a}_n \quad \textit{and} \quad \mathbf{a}_n = \mathbf{L}\mathbf{c}_n,$$

where $\theta/2$ is the population-scaled mutation rate. Furthermore, this allows $\boldsymbol{\tau}_n$ to be computed in $O(n^3)$ time.

Computing the matrix $\mathbf{L}$ in *Theorem 1* is costly. For time-homogeneous coalescents, it is possible to compute $\mathbf{a}_n$ directly, resulting in the following corollary:

Corollary 1. *In the same setting as Theorem 1, if ζ is a constant function, then $\boldsymbol{\tau}_n$ can be computed in $O(n^2)$ time.*

In what follows, *Lemmas 1* and *2* provide formulas to compute the universal matrix $\mathbf{A}$, while *Lemmas 3* and *4* provide formulas to compute $\mathbf{L}$, which is related to the spectral decomposition of the rate matrix $\mathbf{Q}$. The expected first

coalescence times $\mathbf{c}_n = (c_{2,2}, \ldots, c_{n,n})^T$ can be computed as (Polanski and Kimmel 2003; Bhaskar *et al.*, 2015)

$$c_{k,k} = \int_0^\infty \mathbb{P}\{\text{time of first coalescence for } k \text{ individuals} > t\}dt = \int_0^\infty e^{(\mathbf{Q})_{kk}\int_0^t (1/\zeta(s))ds} dt.$$

Note that since $\mathbf{A}$ and $\mathbf{L}$ do not depend on ζ, the SFS depends on time and the inhomogeneity of the coalescent process only through the first coalescence times $\mathbf{c}_n$.

Lemma 1. *Let $\boldsymbol{\gamma}_n := (\tau_{2,1}, \tau_{3,2}, \ldots, \tau_{n,n-1})^T$ denote the antisingleton entries (i.e., entries where exactly one individual has the ancestral allele and all other individuals have the derived allele) of the SFS for samples of sizes $2, \ldots, n$. Then,*

$$\boldsymbol{\tau}_n = \mathbf{B}\boldsymbol{\gamma}_n,$$

where the entries of $\mathbf{B} \in \mathbb{R}^{n-1\times n-1}$ are given by

$$(\mathbf{B})_{ij} = \begin{cases} (-1)^{i-j}\dfrac{1}{j+1}\dbinom{n-i-1}{j-i}\dbinom{n}{i}, & i \le j, \\ 0, & i > j. \end{cases}$$

Proof. We use induction to show that

$$\tau_{n,i} = \sum_{j=i}^{n-1} (-1)^{i-j}\frac{1}{j+1}\binom{n-i-1}{j-i}\binom{n}{i}\tau_{j+1,j}. \tag{1}$$

Using exchangeability and a subsampling argument similar to that of Kamm *et al.* (2015, lemma 2), we obtain, for $k > l+1$,

$$\tau_{k-1,l} = \frac{l+1}{k}\tau_{k,l+1} + \frac{k-l}{k}\tau_{k,l}, \tag{2}$$

which follows from removing an individual uniformly at random from a sample of size k. Now, define the *level* of $\tau_{n,i}$ as $n-i$ and note that (1) holds for level 1, *i.e.*, for $\tau_{l,l-1}$ on the left-hand side. Assume that (1) holds for level $n-i-1$. Then,

$$\begin{aligned}
\tau_{n,i} &= \frac{n}{n-i}\tau_{n-1,i} - \frac{i+1}{n-i}\tau_{n,i+1} \\
&= \frac{n}{n-i}\left[\sum_{j=i}^{n-2}(-1)^{i-j}\frac{1}{j+1}\binom{n-i-2}{j-i}\binom{n-1}{i}\tau_{j+1,j}\right] \\
&\quad - \frac{i+1}{n-i}\left[\sum_{j=i+1}^{n-1}(-1)^{i+1-j}\frac{1}{j+1}\binom{n-i-2}{j-i-1}\binom{n}{i+1}\tau_{j+1,j}\right] \\
&= \binom{n}{i}\left\{\frac{1}{i+1}\tau_{i+1,i} + (-1)^{n-1-i}\frac{1}{n}\tau_{n,n-1}\right. \\
&\quad \left. + \sum_{j=i+1}^{n-2}(-1)^{i-j}\frac{1}{j+1}\left[\binom{n-i-2}{j-i} + \binom{n-i-2}{j-i-1}\right]\tau_{j+1,j}\right\} \\
&= \binom{n}{i}\sum_{j=i}^{n-1}(-1)^{j-i}\frac{1}{j+1}\binom{n-i-1}{j-i}\tau_{j+1,j},
\end{aligned}$$

where the first equality holds by the recursion (2) and the second equality holds by the inductive hypothesis, by noting that $\tau_{n-1,i}$ and $\tau_{n,i+1}$ are both one level below $\tau_{n,i}$.

□

The following lemma relates $\boldsymbol{\gamma}_n$ to $\mathbf{a}_n$:

Lemma 2. *Let γ_n, $\mathbf{a}_n$, and θ be defined as above. Then,*

$$\boldsymbol{\gamma}_n = \frac{\theta}{2}\mathbf{C}\mathbf{a}_n,$$

where $\mathbf{C} \in \mathbb{R}^{n-1\times n-1}$ is bidiagonal with $(\mathbf{C})_{k,k-1} = -(k+1)$ and $(\mathbf{C})_{kk} = k+1$ for $k \in \{2, \ldots, n-1\}$, and $(\mathbf{C})_{11} = 2$.

Proof. As in the *Proof* of *Lemma 1*, we employ a subsampling argument. Consider a sample of size $k+1$. The only way that a subsample of size k can have a different time to the most recent common ancestor is if the removed individual is a singleton after all of the other lineages have coalesced. The probability that we remove that singleton to form our subsample is $1/(k+1)$. Then, the expected amount of time during which there is one singleton and all of the other individuals have coalesced scaled by the mutation rate is exactly the antisingleton entry. Thus,

$$\frac{1}{k+1}\tau_{k+1,k} = \frac{\theta}{2}\left(\mathbb{E}T_{k+1}^{\text{MRCA}} - \mathbb{E}T_k^{\text{MRCA}}\right)$$

for $k > 1$. When $k = 1$, there are only two lineages, so the total branch length is the antisingleton entry. Thus, $\tau_{2,1} = (\theta/2)2\mathbb{E}T_2^{\text{MRCA}}$. Rewriting this as a matrix equation for $k \in \{1, \ldots, n-1\}$ completes the *Proof*.

□

By combining *Lemmas 1* and *2*, we obtain the universal matrix $\mathbf{A} = \mathbf{BC}$. We now show how to compute the Ξ-dependent matrix $\mathbf{L}$. First, we establish the following result on the decomposition of the rate matrix $\mathbf{Q}$; this result was also obtained by Möhle and Pitters (2014, equation 2.3) for the Bolthausen–Sznitman coalescent.

Lemma 3. *Fix an arbitrary Ξ-coalescent with $\lambda_i \neq \lambda_j$ for $i \neq j$, where $\lambda_i := \sum_{k=1}^{i-1}(\mathbf{Q})_{ik} = -(\mathbf{Q})_{ii}$. Let $\mathbf{Q} \in \mathbb{R}^{n\times n}$ denote the rate matrix of the ancestral process corresponding to $\Xi(d\mathbf{x})$ (that is, the process counting the number of extant lineages at time t). Then,*

$$\mathbf{Q} = \mathbf{U}\mathbf{E}\mathbf{U}^{-1},$$

where $(\mathbf{E})_{ij} = \delta_{ij}(\mathbf{Q})_{ii}$, with δ_{ij} being the Kronecker delta that equals 1 if $i = j$ and 0 otherwise, and

$$(\mathbf{U})_{ij} = \begin{cases} 1, & i = j, \\ \dfrac{1}{\lambda_i - \lambda_j}\displaystyle\sum_{k=j}^{i-1}(\mathbf{Q})_{ik}(\mathbf{U})_{kj}, & i > j, \\ 0, & \text{otherwise}. \end{cases}$$

Proof. By the construction of $\mathbf{U}$,

$$(\mathbf{U})_{ij}(\mathbf{Q})_{jj} = \sum_{k=j}^{i} (\mathbf{Q})_{ik}(\mathbf{U})_{kj},$$

which implies that $\mathbf{UE} = \mathbf{QU}$. Then, since $\mathbf{U}$ is triangular and has strictly positive diagonal entries, it is invertible. Therefore, $\mathbf{Q} = \mathbf{UEU}^{-1}$.

□

The following result relates $\mathbf{a}_n := (\mathbb{E}T_2^{\text{MRCA}}, \ldots, \mathbb{E}T_n^{\text{MRCA}})^T$ and $\mathbf{c}_n = (c_{2,2}, \ldots, c_{n,n})^T$:

Lemma 4. *Let $\mathbf{a}_n$ and $\mathbf{c}_n$ be defined as above. Fix an arbitrary Ξ measure and a strictly positive function ζ. Now consider a time-inhomogeneous coalescent governed by $\Xi(d\mathbf{x})/\zeta(t)$. If $c_{k,k} < \infty$, for $2 \le k \le n$, then*

$$\mathbf{a}_n = -(\mathbf{UD})_{2:n,2:n}\mathbf{c}_n,$$

where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is the diagonal matrix $\operatorname{diag}([\mathbf{U}^{-1}]_{\cdot,1})$, with $[\mathbf{U}^{-1}]_{\cdot,1}$ denoting the first column of $\mathbf{U}^{-1}$, and $(\mathbf{UD})_{2:n,2:n}$ denotes the submatrix of $\mathbf{UD}$ in rows and columns 2 through n.

Proof. Note that $\mathbb{E}T_k^{\text{MRCA}} = \int_0^\infty \mathbb{P}\{T_k^{\text{MRCA}} > t\}dt$. Therefore,

$$\begin{aligned}\mathbb{E}T_k^{\text{MRCA}} &= \int_0^\infty \mathbb{P}\{T_k^{\text{MRCA}} > t\}dt = \int_0^\infty \sum_{l=2}^{k}\left[e^{\mathbf{Q}\int_0^t (1/\zeta(s))ds}\right]_{kl} dt \\ &= \int_0^\infty \sum_{l=2}^{n}\left[e^{\mathbf{Q}\int_0^t (1/\zeta(s))ds}\right]_{kl} dt \\ &= \int_0^\infty \sum_{l=2}^{n}\left[\mathbf{U}e^{\mathbf{E}\int_0^t (1/\zeta(s))ds}\mathbf{U}^{-1}\right]_{kl} dt,\end{aligned}$$

where the third equality follows from the fact that $\mathbf{Q}$ is lower triangular and hence so is its exponential. Now, since $\mathbf{U}$ is lower triangular, its inverse is as well. Therefore, we may ignore the value of $[e^{\mathbf{E}\int_0^t (1/\zeta(s))ds}]_{1,1}$. Letting $\mathbf{F}(t) := e^{\mathbf{E}\int_0^t (1/\zeta(s))ds}$ but with $\mathbf{F}_{1,1}(t) := 0$, note that $\int_0^\infty \mathbf{F}(t)dt = \operatorname{diag}(0, \mathbf{c}_n)$. Then we have

$$\begin{aligned}\mathbb{E}T_k^{\text{MRCA}} &= \int_0^\infty \sum_{l=2}^{n}\left[\mathbf{UF}(t)\mathbf{U}^{-1}\right]_{kl} dt \\ &= \sum_{l=2}^{n}[\mathbf{U}\operatorname{diag}(0, \mathbf{c}_n)\mathbf{U}^{-1}]_{kl}.\end{aligned}$$

Now, note that $(\mathbf{U})_{i,1} = 1$ for all i by *Lemma 3* and induction. This implies $\sum_{l=1}^n [\mathbf{U}^{-1}]_{il} = \delta_{i1}$, or $\sum_{l=2}^n [\mathbf{U}^{-1}]_{il} = \delta_{i1} - [\mathbf{U}^{-1}]_{i1}$. Using this identity, we can rewrite the above expression for $\mathbb{E}T_k^{\text{MRCA}}$ as

$$\mathbb{E}T_k^{\text{MRCA}} = -\sum_{j=2}^{n}\left[(\mathbf{UD})_{2:n,2:n}\right]_{k-1,j} c_{j,j},$$

where $\mathbf{D} = \operatorname{diag}([\mathbf{U}^{-1}]_{\cdot,1})$. Collecting these equations over $k \in \{2, \ldots, n\}$ in matrix form leads to the desired result.

□

Using *Lemma 4*, we now see that the matrix $\mathbf{L}$ from *Theorem 1* is simply $-(\mathbf{UD})_{2:n,2:n}$. *Lemma 3* provides a recursion to compute $\mathbf{U}$, and $\mathbf{D}$ may be computed by noting that $(\mathbf{U}^{-1})_{11} = 1$ and then since $\mathbf{UU}^{-1} = \mathbf{I}$, we have

$$\mathbf{U}_{i1}^{-1} = -\sum_{j=1}^{i-1}(\mathbf{U})_{ij}(\mathbf{U}^{-1})_{j1}.$$

Proof of Theorem 1. Combining *Lemmas 1*, *2*, and *4*, we obtain the equations in *Theorem 1*. For the runtime, note that each of the $O(n^2)$ entries of $\mathbf{U}$ requires $O(n)$ computations, and so computing $\mathbf{U}$ is $O(n^3)$. The matrices composing $\mathbf{A}$ are known in closed form, however, and constructing $\mathbf{D}$ requires filling only $O(n)$ entries, each requiring $O(n)$ computations for a total of $O(n^2)$. To then obtain the SFS from $\mathbf{c}_n$ simply requires iterated matrix vector products taking $O(n^2)$ time. The overall procedure thus requires $O(n^3)$.

□

Lemma 5. *For coalescents of the form $\Xi(d\mathbf{x})/\zeta(t)$ where ζ is a constant function, $\mathbf{a}_n$ can be computed recursively from $\mathbf{c}_n$ and $\mathbf{Q}$ as follows*:

$$\begin{aligned}\mathbb{E}T_2^{\text{TMRCA}} &= c_{2,2} \\ \mathbb{E}T_k^{\text{TMRCA}} &= c_{k,k} + \sum_{l=2}^{k-1}\frac{(\mathbf{Q})_{kl}}{\lambda_k}\mathbb{E}T_l^{\text{TMRCA}}, \quad \text{for } k > 2.\end{aligned}$$

Proof. The formulas follow immediately from the homogeneity of the process, recursing on the number of individuals, and noting that the probability that the first coalescence event for a sample of size k results in k lineages merging down to l lineages is $(\mathbf{Q})_{kl}/\lambda_k$.

□

Proof of Corollary 1. Use *Lemma 5* to compute $\mathbf{a}_n$ in $O(n^2)$ time. Then, $\boldsymbol{\tau}_n = \mathbf{Aa}_n$ by *Theorem 1*, which also takes $O(n^2)$ time to compute.

□

Remark 1. *Other than computing $\mathbf{U}$, the algorithm presented in Theorem 1 is $O(n^2)$. Thus, for the Bolthausen–Sznitman coalescent (Bolthausen and Sznitman 1998) or Kingman's coalescent, where $\mathbf{U}$ is known in closed form (Möhle and Pitters 2014, theorem 1.1 and appendix), the SFS can be computed in $O(n^2)$ time even for nonconstant ζ.*

Remark 2. *The above results can easily be extended to a coalescent where both ζ and $\Xi(d\mathbf{x})$ depend on t, as long as $\Xi(d\mathbf{x})$ is piecewise constant. For example, in the recent past the population may evolve according to a β-coalescent, whereas for t greater than some t_0 the population may evolve according to Kingman's coalescent. By setting ζ appropriately in Theorem 1, one may obtain a "truncated SFS" (Kamm et al. 2015) for each different $\Xi(d\mathbf{x})$. Then, using the truncated SFS for each epoch*

and the same machinery as in Kamm et al. (2015), *one may compute the full SFS. The same techniques also allow one to consider multiple populations, with each population perhaps evolving according to its own Ξ measure.*

Numerical Results

We implemented *Theorem 1* and *Corollary 1* in Mathematica, and the notebook is available upon request. We can compute the SFS for an arbitrary coalescent for a sample of size $n = 100$ in $\sim$1 sec and a sample of size $n = 300$ in a matter of minutes on a laptop computer, which is orders of magnitude faster than the $>$1 hr reported for a sample size of $n = 100$ using the current state-of-the-art method (Blath *et al.* 2015). Furthermore, Blath *et al.* (2015) consider only specific Ξ measures where the number of simultaneous multiple mergers is restricted. Our method has the same runtime for all Ξ measures (after computing the rate matrix and the vector of first coalescence times). See Figure 1 for runtime *vs.* sample size. Furthermore, as noted above, if the spectral decomposition of the rate matrix $\mathbf{Q}$ is known, then the algorithm is $O(n^2)$. We also present runtimes for the Bolthausen–Sznitman coalescent [which has a closed-form solution for the spectral decomposition (Möhle and Pitters 2014)] in Figure 1.

As long as the rate matrix $\mathbf{Q}$ of the ancestral process can be found exactly, our method is numerically stable. This is the case for popular Λ-coalescents such as point-mass coalescents and β-coalescents, as well as point mass Ξ-coalescents. If the rate matrix must be evaluated numerically, however, high-precision computation may be needed to avoid potential numerical problems due to catastrophic cancellation.

Using simulations, we now investigate the variation in the empirical SFS across independent realizations of the coalescent process and study the ability to infer the underlying model, using the empirical SFS. We consider three different ζ's, illustrated in Figure 2. Due to the association with population sizes in the case of Kingman's coalescent, we refer to ζ as the history or population size history. However, we caution that depending on the finite population size model, ζ may not represent the population size, but some other biologically relevant parameter. We consider a constant size history, a bottleneck history that undergoes a temporary 10-fold size reduction, and a growth history with repeated population doublings. For each ζ, we consider $\beta(2-\alpha, \alpha)$-coalescents with $\alpha \in \{1, 1.5, 2\}$. Note that $\alpha = 1$ corresponds to the Bolthausen–Sznitman coalescent, while $\alpha = 2$ corresponds to the Kingman coalescent. For each of the nine distinct values of (ζ, α), we simulated $m = 1000$ independent trees with $n = 20$ leaves.

In Figure 3, we examine the observed variation in branch lengths across independent realizations of the coalescent process, from which we can deduce the variation in the observed SFS. Specifically, assume that each tree sampled from the coalescent process has the same mutation rate, and, without loss of generality, assume that time has been scaled such that the mutation rate is 1. Let $\tilde{\tau}_{n,k}$ be the sum of branch lengths with k leaves and recall that $\hat{\tau}_{n,k}$ is the kth entry of the empirical SFS on the n observed individuals. Then, $\mathbb{P}(\tilde{\boldsymbol{\tau}}_n | \tilde{\boldsymbol{\tau}}_n) \sim \text{Poisson}(\tilde{\boldsymbol{\tau}}_n)$, and $\mathbb{E}[\tilde{\boldsymbol{\tau}}_n] = \boldsymbol{\tau}_n$. In Figure 3, we plot $\tilde{\tau}_{n,k}$ for each simulated tree, as well as its expected value $\mu_{n,k}$. Defining $\sigma^2_{n,k} := \text{Var}(\tilde{\tau}_{n,k})$ for this case of $m = 1$, we also plot an estimate of the standard deviation $\hat{\sigma}_{n,k} = \sqrt{\hat{E}[\tilde{\tau}^2_{n,k}] - \tau^2_{n,k}}$, where $\hat{E}$ is the empirical expectation. Now, if we sum the branch lengths and mutations over m independent trees [so then $\mathbb{E}[\tilde{\tau}_{n,k}] = m\mu_{n,k}$ and $\text{Var}(\tilde{\tau}_{n,k}) = m\sigma^2_{n,k}$], then $\mu_{n,k}$ and $\sigma^2_{n,k}$ describe the limiting behavior of both $\tilde{\tau}_{n,k}$ and $\hat{\tau}_{n,k}$ as $m \to \infty$: by the central limit theorem, $(1/\sqrt{m})\ (\tilde{\tau}_{n,k} - \tau_{n,k}) \to_d \mathcal{N}(0, \sigma^2_{n,k})$ and $(1/\sqrt{m})(\hat{\tau}_{n,k} - \tau_{n,k}) \to_d \mathcal{N}(0, \sigma^2_{n,k} + \mu_{n,k})$.

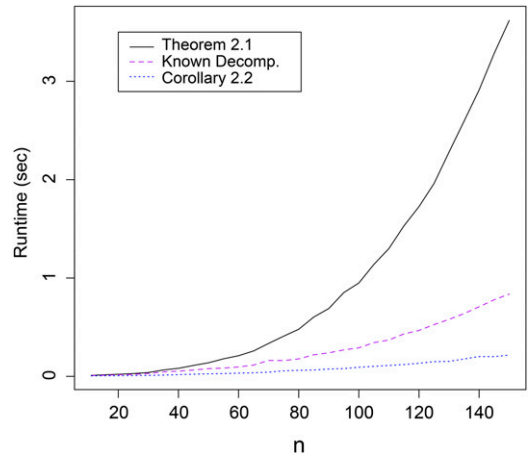

Figure 1 Runtime result (in seconds). *Theorem 1* was used to compute the SFS for the time-homogeneous Bolthausen–Sznitman coalescent. The solid line uses *Lemma 3* to compute the spectral decomposition of $\mathbf{Q}$ resulting in a cubic runtime. The dashed line uses the closed-form representation of the spectral decomposition of the Bolthausen–Sznitman coalescent (Möhle and Pitters 2014, theorem 1.1) to compute the SFS in quadratic time. The dotted line uses *Corollary 1*, which is also quadratic.

A recent inconsistency result (Koskela *et al.* 2015, theorem 1) shows that a Λ measure cannot be inferred from a single tree ($m = 1$), even as $n \to \infty$. Indeed, we see in Figure 3 that the branch lengths $\tilde{\tau}_{n,k}$ of a single tree can deviate substantially from $\tau_{n,k}$. For most k (say, $k \geq 5$), typically $\tilde{\tau}_{n,k} = 0$ or $\tilde{\tau}_{n,k} \gg \mu_{n,k}$, given a single tree. That is, for a single tree, branches subtending more than a few leaves are either not observed or much larger than the expected branch length. However, smaller k (especially the singletons, $k = 1$) have smaller relative standard deviation $\sigma_{n,k}/\mu_{n,k}$ and thus will tend to have lower relative error $(\hat{\tau}_{n,k} - \tau_{n,k})/\tau_{n,k} \approx \mathcal{N}(0, (\sigma^2_{n,k} + \mu_{n,k})/m\mu^2_{n,k})$ as m increases.

In the case of Kingman's coalescent, ζ is inferred by minimizing the Kullback-Leibler (KL) divergence between a normalized version of the empirical SFS and a normalized version of the expected SFS (*e.g.*, Bhaskar *et al.* 2015, equation 10). Recall that the KL divergence, $D_{\text{KL}}(\Phi_1 \| \Phi_2)$ between two discrete probability distributions, Φ_1 and Φ_2 is $\sum_i \Phi_{1(i)} \log(\Phi_{1(i)} \Phi_{2(i)})$ where $0\log(0)$ is defined to be 0. We

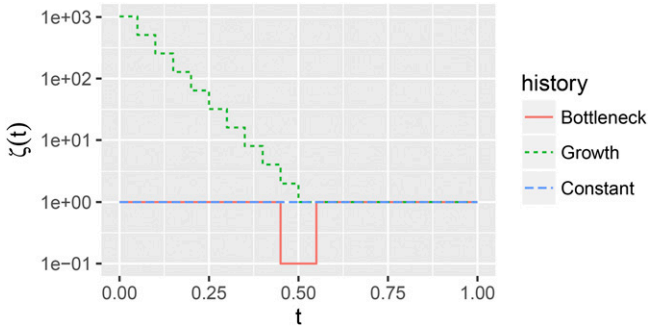

Figure 2 $\zeta(t)$ for three demographic scenarios: a constant size history, a bottleneck history that undergoes a temporary 10-fold size reduction, and a growth history with repeated population doublings. Note that the y-axis is stretched by $y \mapsto \log(y)$.

investigate how KL divergence behaves as a function of the number m of independent trees simulated in the case of Λ-coalescents. Let $\boldsymbol{\tau}_n^{(\zeta,\alpha)}$ be the expected SFS under model $(\zeta(t),\alpha)$ and $\tilde{\boldsymbol{\tau}}_n^{(\zeta,\alpha)}(m)$ the corresponding branch lengths summed over the first m simulated trees. Define $P^{(\zeta,\alpha)}(k) \propto \tau_{n,k}^{(\zeta,\alpha)}$ as the true probability distribution of derived alleles under scenario $(\zeta(t),\alpha)$ and $\tilde{P}_m^{(\zeta,\alpha)}(k) \propto \tilde{\tau}_{n,k}^{(\zeta,\alpha)}(m)$ as the conditional distribution of derived alleles, given the first m trees simulated under $(\zeta(t),\alpha)$. In Figure 4, we plot the KL divergence $D_{\mathrm{KL}}(\tilde{P}_m^{(\zeta_1,\alpha_1)} \| P^{(\zeta_2,\alpha_2)})$ as a function of m, for every $(\zeta_1(t),\zeta_2(t),\alpha_1,\alpha_2)$ considered above (that is, ζ is constant, bottleneck, or growth, and α is 1, 1.5, or 2). In this case, we see that minimizing D_{KL} identifies the true scenario $(\zeta_1(t),\alpha_1(t)) = (\zeta_2(t),\alpha_2(t))$ with access to only a moderate number of independent trees (between 10 and 100).

Figure 4 is encouraging, as not too many independent trees are needed to distinguish between the different scenarios $(\zeta(t),\alpha)$. Unfortunately, in some cases it may be impossible to even sample two independent trees (J. Koskela, personal communication). For example, in the model of Birkner *et al.* (2013b), a multiple-merger event happens over a single "generation," which can cause the multiple merger to affect unlinked sites, resulting in correlated coalescence times. However, in other models, multiple-merger events may affect the genome only locally, and thus trees from unlinked sites are independent. For example, in the selective sweep model of Durrett and Schweinsberg (2005), multiple mergers are caused by selective sweeps taking place over $O(\log(N))$ "generations," and a site experiences a multiple merger if $r_N \log(2N)/s_N = O(1)$, where N, s_N, r_N respectively parameterize the population size, selection strength, and recombination distance to the selected site. Thus, the independence of unlinked trees is not necessarily determined by the Λ or Ξ measure itself, but instead by the prelimiting model.

Identifiability Results

Before attempting to infer ζ or Ξ in practice, it is important to know whether such inference is possible using the SFS. For instance, when inferring ζ, if two different functions ζ_1 and ζ_2 produce the same SFS, then it is impossible to distinguish between the two using only the SFS. In such a case, we say that ζ is not identifiable. For Kingman's coalescent if one allows ζ to be an arbitrary positive function that produces a finite SFS, then ζ is not identifiable (Myers *et al.* 2008). ζ is identifiable in the case of Kingman's coalescent, however, if one restricts ζ to be from a set of biologically realistic functions (technically, a set of functions with only a finite number of oscillations) (Bhaskar and Song 2014, theorem 11). We show that a similar result holds for all coalescents of the form $\Xi(d\mathbf{x})/\zeta(t)$ where $\Xi(d\mathbf{x})$ is fixed.

In general it is impossible to infer Ξ from the SFS if Ξ is not restricted. There has been some interest, however, in the case of distinguishing between a subset of Λ-coalescents (Eldon *et al.* 2015). We prove some results about the identifiability of the measure for various subsets of Λ measures when ζ is a constant function. We also consider the question posed by Eldon *et al.* (2015) of whether the SFS can distinguish between exponential growth under Kingman's coalescent and a class of Λ-coalescents with constant ζ, and we show that indeed it is possible to distinguish between these cases with a surprisingly small number of samples. We note that our identifiability results require knowledge of the exact expected SFS, whereas Eldon *et al.* (2015) focus on the case where the expected SFS is approximated using an empirical SFS, which is what occurs in practice.

Throughout this section we assume that one has the exact expected SFS (*i.e.*, the object computed by *Theorem 1*).

Identifiability of ζ for fixed Ξ measure

Before proceeding to the results and proofs, we first introduce some notation. Let $\mathcal{M}_K(\mathcal{F})$ denote the set of piecewise defined functions with at most K pieces made from some function family $\mathcal{F}$. Furthermore, let $\mathcal{S}(\mathcal{F})$ denote the sign-change complexity of $\mathcal{F}$. Informally, $\mathcal{S}(\mathcal{F})$ is the supremum of the number of times $f_1 - f_2$ crosses 0 over functions $f_1, f_2 \in \mathcal{F}$, which is related to the number of oscillations each $f \in \mathcal{F}$ is allowed to have [see Bhaskar and Song 2014, definition 4, for a formal definition of $\mathcal{S}(\mathcal{F})$]. We will also write ψ_n^{Ξ} for the number of 0 entries in $[\mathbf{U}^{-1}]_{\cdot,1}$ in the spectral decomposition of $\mathbf{Q}$ for a coalescent on n individuals governed by $\Xi(d\mathbf{x})$. Furthermore, denote by χ the space of Ξ measures such that $(\mathbf{Q})_{k,k-1} > 0$ for all k. That is, χ is the set of Ξ measures where for any sample size there is positive probability of a single pairwise merger. If we are considering only Λ-coalescents, then χ contains all Λ measures except for δ_1, the star coalescent. We now present our main identifiability results and a conjectured bound on ψ_n^{Ξ}.

Our main result on the identifiability of ζ is the following theorem.

Theorem 2. *For an arbitrary Ξ-coalescent governed by the measure $\Xi(d\mathbf{x})/\zeta(t)$ where $\Xi \in \chi$ is fixed, suppose $\mathcal{S}(\mathcal{F}) < \infty$ and $n \geq 2K + (2K-1)\mathcal{S}(\mathcal{F}) + \psi_n^{\Xi}$. Then for each expected SFS τ_n there exists a unique $\zeta \in \mathcal{M}_K(\mathcal{F})$ consistent with τ_n.*

First, note that in the case of Kingman's coalescent, $\psi_n^{\delta_0} = 0$ for all n, and so in some sense, Kingman's coalescent is

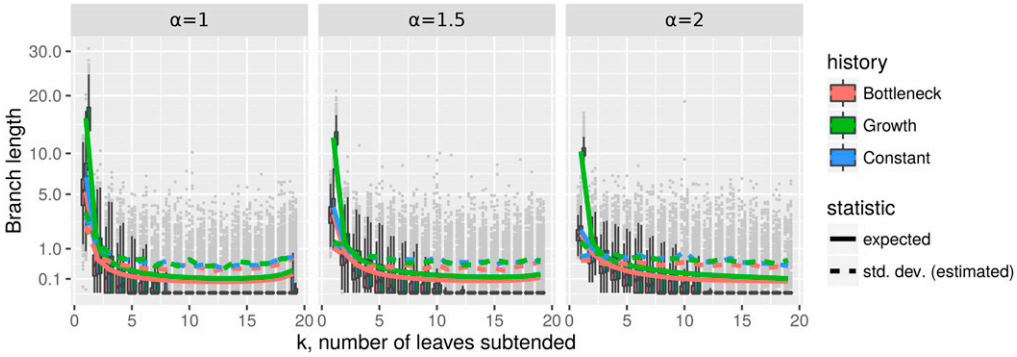

Figure 3 The distribution of the branch length subtending k leaves, for random trees under a $\beta(2-\alpha,\alpha)$-coalescent and $n=20$. The solid line is the expected value from *Theorem 1*. We simulated 1000 independent trees per scenario and their branch length results are shown as gray dots and box plots; the dashed line denotes the estimated standard deviation of the distribution. Note the y-axis is stretched by $y\mapsto\sqrt{y}$. The mean and standard deviation give the limiting behavior of $\hat{\tau}_{n,k}$ for many independent trees, under the central limit theorem. For most k (say, $k\geq 5$), the branch length is usually 0 and has high variance relative to the mean. Thus $\hat{\tau}_{n,k}$ will tend to have higher relative accuracy for the smaller entries k.

optimal in terms of the number of samples needed to ensure that a certain model space is identifiable. For the Bolthausen–Sznitman coalescent, $\psi_n^1=0$ for all n, which follows from the spectral decomposition (Möhle and Pitters 2014, theorem 1.1). For the point mass Λ-coalescent with mass at $1/2$, for $n\geq 5$, all odd entries of $[\mathbf{U}^{-1}]_{\cdot,1}$ are 0 and so $\psi_n^{\delta_{1/2}}>0$ for $n\geq 5$, thus implying that larger samples (relative to Kingman's coalescent or the Bolthausen–Sznitman coalescent) are needed for this coalescent to ensure that a given model space is identifiable. We suspect that $\Lambda(dx)=\delta_{1/2}$ is the worst case among all Ξ-coalescents in χ for identifiability, resulting in the following conjecture:

Conjecture 1. *For all $\Xi\in\chi$ and $n\geq 3$, $\psi_n^{\Xi}\leq(n-1)/2-1$.*

If this conjecture is true, then the bound on the sample size needed to have identifiability in *Theorem 2* can be simplified to $n\geq 2[2K+(2K-1)\mathscr{S}(\mathcal{F})]$

Identifiability of the Λ measure for a constant ζ

We also have the following results for Λ-coalescents about the identifiability of the Λ measure.

Theorem 3. *Consider the set of point-mass Λ-coalescents: $\{\delta_z : z\in[0,1]\}$ If Λ is restricted to be in this set and $n\geq 3$, then the expected SFS τ_n uniquely determines Λ.*

Theorem 4. *Consider the set of β-coalescents: $\{\mathcal{L}(X) : X\sim\beta(2-\alpha,\alpha),\alpha\in[1,2)\}$ If Λ is restricted to be in this set and $n\geq 3$, then the expected SFS τ_n uniquely determines Λ.*

Theorem 5. *Consider the set of coalescents $\{\delta_0/ae^{-bt}: a,b>0\}\cup\{\delta_z : z\in[0,1]\}\cup\{\mathcal{L}(X) : X\sim\beta(2-\alpha,\alpha),\alpha\in[1,2)\}$, that is, Kingman's coalescent with exponential growth, point-mass coalescents, or β-coalescents. If Λ is restricted to be in this set and $n\geq 4$, then the expected SFS τ_n uniquely determines Λ.*

Theorem 5 gives a positive theoretical answer to the question of whether the SFS can distinguish between exponential growth and multiple-merger coalescents. Using the techniques presented below, it is straightforward to obtain similar results for other subsets of Λ-coalescents.

Proofs of the identifiability results

The following lemma is used in proving the theorems in this section and may be of independent interest, as it shows that given the SFS for n individuals one can compute the expected time to most recent common ancestor for sample sizes $2,\ldots,n$ or vice versa.

Lemma 6. *For all Λ- and Ξ-coalescents, there is a bijection between the expected SFS τ_n and the expected times $\mathbf{a}_n$ to the most recent common ancestor.*

Proof. Combine *Lemmas 1* and *2* to see that $\boldsymbol{\tau}_n=\mathbf{BCa}_n$, with $\mathbf{B}$ and $\mathbf{C}$ being universal. Then, since $\mathbf{B}$ is upper triangular and all of its diagonal entries are nonzero, it is invertible. Furthermore, since $\mathbf{C}$ is bidiagonal and the diagonal entries are all nonzero, it is also invertible. Therefore, $\mathbf{BC}$ is invertible and since $\boldsymbol{\tau}_n$ and $\mathbf{a}_n$ are related through an invertible matrix, the transformation is bijective.

□

To prove *Theorem 2* we use the following lemma.

Lemma 7. *Let $\lambda_k=-(\mathbf{Q}_{kk})$. For all $\Xi\in\chi$ and all Λ other than $\Lambda(dx)=\delta_1(dx)$ (i.e., the star coalescent), the sequence $(\lambda_k)_{k\geq 2}$ is strictly increasing.*

Proof. Consider a sample of size $k+1$ and a subsample of size k. Without loss of generality, assume individual $k+1$ is removed to produce the subsample. The time to the first event is the same for both samples unless the first event involves only individual $k+1$ and one lineage from $\{1,\ldots,k\}$. That is, the total rate when there are $k+1$ lineages is equal to the total rate when there are k lineages plus k times the rate at which exactly a particular pair of individuals coalesce. Formally,

$$\lambda_{k+1}=\lambda_k+\frac{k}{\binom{k+1}{2}}(\mathbf{Q})_{k+1,k}.$$

By assumption, $(\mathbf{Q})_{k+1,k}>0$, and so the total rates must be strictly increasing.

□

We now prove *Theorem 2*. Our proof relies heavily on the proof of the corresponding result for Kingman's coalescent (Bhaskar and Song 2014, theorem 11). We essentially show that this setting satisfies the same hypotheses as the Kingman's coalescent case and then use that result to complete our proof.

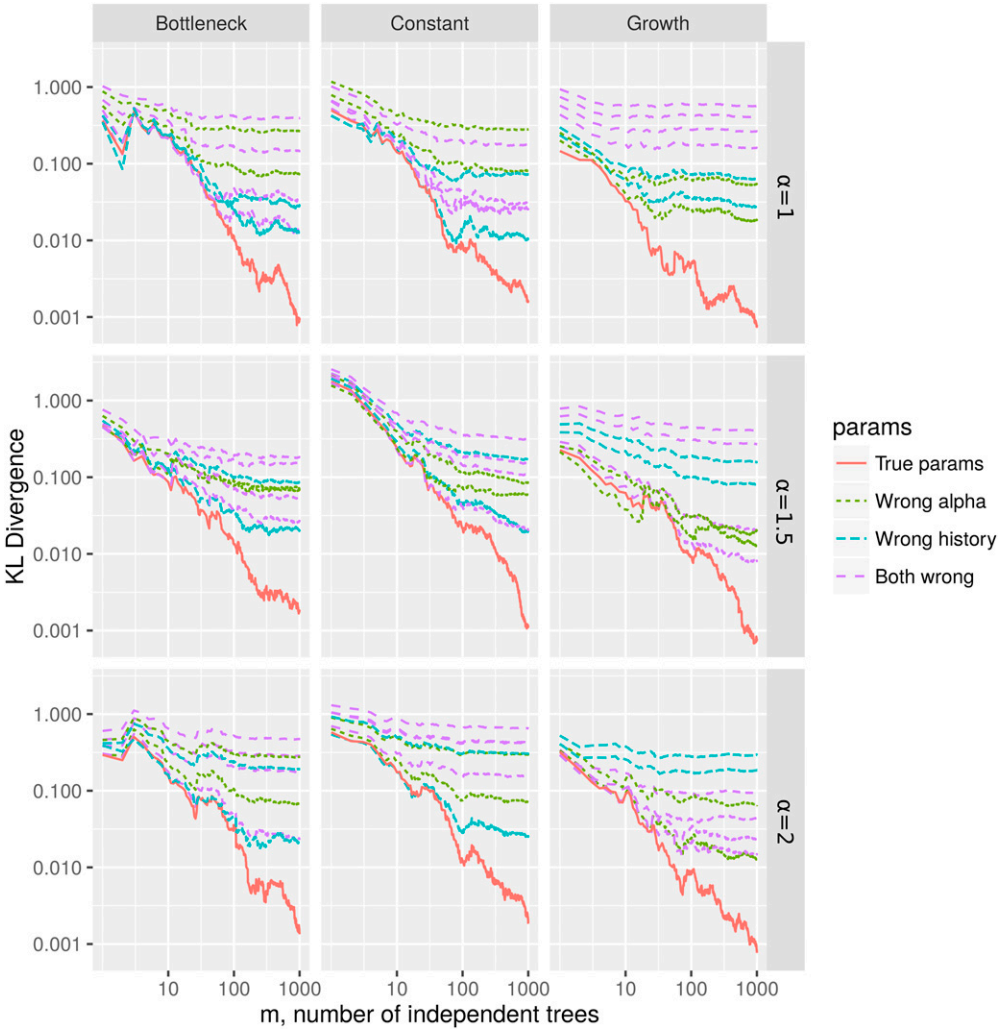

Figure 4 The KL divergence $D_{\mathrm{KL}}(\tilde{P}_m^{(\zeta_1,\alpha_1)} \| P^{(\zeta_2,\alpha_2)})$, where $P^{(\zeta_2,\alpha_2)}(k) \propto \tau_{n,k}^{(\zeta_2,\alpha_2)}$ is the distribution of derived alleles under scenario $(\zeta_2(t),\alpha_2)$, and $\tilde{P}_m^{(\zeta_1,\alpha_1)}(k) \propto \tilde{\tau}_{n,k}^{(\zeta_1,\alpha_1)}$ (m) is the conditional distribution of derived alleles, given the first m trees simulated under $(\zeta_1(t),\alpha_1)$ and a mutation hitting one of those trees. For m large enough, D_{KL} is minimized by the true parameters; *i.e.*, $(\zeta_1(t),\alpha_1(t)) = (\zeta_2(t),\alpha_2(t))$. D_{KL} can typically discriminate the true scenario for $m = 100$ trees. For $m = 10$ trees, D_{KL} is often, but not always, minimized by the true scenario.

Proof of Theorem 2. By *Lemma* 6, the SFS is uniquely determined by $\mathbf{a}_n$. Then, furthermore, note that from *Lemma 3* the matrix $\mathbf{U}$ is invertible since it is triangular with all nonzero entries along the diagonal. Then, by the same argument as in Bhaskar and Song (2014, equation 12), we know that if the model space is not identifiable, then for each k not corresponding to a zero in $\mathbf{D}$ (contributing to ψ_n^{Ξ}), λ_k must be the root of the Laplace transform of two different functions in the model space. By *Lemma 7*, these are all distinct, resulting in $n - \psi_n^{\Xi}$ roots. Then, by taking $n - \psi_n^{\Xi}$ sufficiently large, we obtain a contradiction via the generalized version of Descarte's rule of signs (Bhaskar and Song 2014, theorem 4) and the theorem is proved.

□

We now prove *Theorems 3*, *4*, and *5*. The idea is to explicitly calculate the $\mathbb{E}T_k^{\mathrm{MRCA}}$ for the first few k for each allowed Λ measure and then use *Lemma 6* to show that if Λ is uniquely determined by the first few $\mathbb{E}T_k^{\mathrm{MRCA}}$, then it is uniquely determined by $\boldsymbol{\tau}_n$.

Proof of Theorem 3. $\mathbb{E}T_2^{\mathrm{MRCA}} = 1$ for all Λ in the set of possible Λ's. Consider $\Lambda = \delta_z$. Using *Lemma 5* we see that

$$\mathbb{E}T_3^{\mathrm{MRCA}} = \frac{1}{3-2z} + \frac{3-3z}{3-2z} \cdot 1 = \frac{4-3z}{3-2z}. \tag{3}$$

This is a monotonically decreasing function of $z \in [0,1]$, and so Λ is uniquely determined by $\mathbb{E}T_3^{\mathrm{MRCA}}$. Then, appealing to *Lemma 6*, we see that Λ is uniquely determined by the SFS for $n \geq 3$.

□

Proof of Theorem 4. A calculation similar to (3) gives $\mathbb{E}T_3^{\mathrm{MRCA}} = (2+3\alpha)/(2+2\alpha)$ for $\Lambda = \mathcal{L}(X), X \sim \beta(2-\alpha,\alpha)$, where $\alpha \in [1,2)$. This is a monotonically increasing function of $\alpha \in [1,2)$ and the claim follows from the same argument as in the proof of *Theorem 3*.

□

Proof of Theorem 5. Suppose that two distinct Λ-coalescents within the set of allowed models produce the same expected SFS for $n \geq 4$. Then, by *Lemma 6*, they would have the same values of $\mathbb{E}T_2^{\mathrm{MRCA}}$, $\mathbb{E}T_3^{\mathrm{MRCA}}$, and $\mathbb{E}T_4^{\mathrm{MRCA}}$. By *Theorems 3* and *4*, we know that the Λ measures cannot both be point-mass coalescents or β-coalescents. From Bhaskar and Song (2014,

corollary 8), we also know that the Λ measures cannot both be Kingman's coalescent with different exponential growth parameters. There are thus three cases. They are all straightforward, albeit tedious.

Case 1. One Λ measure is a point-mass coalescent and the other is a β-coalescent. Letting $\mathbb{E}T_2^{\mathrm{MRCA}} = 1$ (without loss of generality), we can explicitly compute $\mathbb{E}T_4^{\mathrm{MRCA}}$ for the point-mass coalescent and the β-coalescent, using the same recursive idea as in the *Proof of Theorem 3*. Let $p_{m,k}$ denote the probability that when there are m lineages, exactly k of them are involved in the next coalescence event. Then, by *Lemma 5* $\mathbb{E}T_4^{\mathrm{MRCA}} = c_{4,4} + p_{4,2}\mathbb{E}T_3^{\mathrm{MRCA}} + p_{4,3}\mathbb{E}T_2^{\mathrm{MRCA}}$. In particular, for the point-mass coalescent δ_z, this implies $\mathbb{E}T_4^{\mathrm{MRCA}} = 5/3 + 1/(2z-3) + (3-2z)/(18-24z+9z^2)$. Now, recalling the expression of $\mathbb{E}T_3^{\mathrm{MRCA}}$ in (3) and letting

$$\mathbb{E}T_3^{\mathrm{MRCA}} = t \tag{4}$$

implies $z = (3t-4)/(2t-3)$. Plugging this into $\mathbb{E}T_4^{\mathrm{MRCA}}$, we see that for the point-mass coalescent,

$$\mathbb{E}T_4^{\mathrm{MRCA}} = \frac{1}{3}\left[6t - 4 - \frac{2t-3}{6+t(3t-8)}\right]. \tag{5}$$

A similar calculation for the β-coalescent shows that

$$\mathbb{E}T_4^{\mathrm{MRCA}} = \frac{1}{3}\left(6t - 2 + \frac{1}{t-2}\right), \tag{6}$$

with $t := \mathbb{E}T_3^{\mathrm{MRCA}}$ under the β-coalescent. Equating (5) and (6) and solving for t results in the solution $t = 1$ or $t = 4/3$. But, if $t = 1$, then we see that $z = 1$, the star coalescent, which corresponds to $\alpha = 0$ for the β-coalescent, which is not in the set of allowed β-coalescents. If $t = 4/3$, we see that $z = 0$, which corresponds to Kingman's coalescent, and $\alpha = 2$ for the β-coalescent, which again is not in the set of allowed β-coalescents. Therefore, a point-mass coalescent and a β-coalescent with $\alpha \in [1,2)$ cannot have the same $\mathbb{E}T_2^{\mathrm{MRCA}}, \mathbb{E}T_3^{\mathrm{MRCA}}$, and $\mathbb{E}T_4^{\mathrm{MRCA}}$ simultaneously.

Case 2. One Λ measure is a point-mass coalescent and the other is Kingman's coalescent with exponential growth. Without loss of generality, assume that $\mathbb{E}T_2^{\mathrm{MRCA}} = 1$ for the point-mass Λ-coalescent. The exponential-growth Kingman's coalescent model considered here has

$$c_{m,m} = -(1/b)e^{\binom{m}{2}\Big/(ab)}\mathrm{Ei}\left[-\binom{m}{2}\Big/(ab)\right], \quad \text{where}$$

$\mathrm{Ei}(x) := -\int_{-x}^{\infty}(e^{-t}/t)dt$ is the exponential integral (Bhaskar *et al.* 2015, supplemental material equation 5). Then, the constraint $\mathbb{E}T_2^{\mathrm{MRCA}} = c_{2,2} = 1$ implies $b = -e^{1/d}\mathrm{Ei}(-1/d)$, where $d := ab$. Furthermore, assuming this constraint and applying *Theorem 1* to Kingman's coalescent, we obtain

$$\mathbb{E}T_3^{\mathrm{MRCA}} = \frac{3}{2} - \frac{c_{3,3}}{2} = \frac{3}{2} - \frac{e^{2/d}\mathrm{Ei}(-3/d)}{2\mathrm{Ei}(-1/d)}, \tag{7}$$

$$\begin{aligned}\mathbb{E}T_4^{\mathrm{MRCA}} &= \frac{9}{5} - c_{3,3} + \frac{c_{4,4}}{5} \\ &= \frac{9}{5} - \frac{e^{2/d}\mathrm{Ei}(-3/d)}{\mathrm{Ei}(-1/d)} + \frac{e^{5/d}\mathrm{Ei}(-6/d)}{5\mathrm{Ei}(-1/d)}.\end{aligned} \tag{8}$$

Now, in addition to $\mathbb{E}T_2^{\mathrm{MRCA}}$, if the two coalescents have the same values of $\mathbb{E}T_3^{\mathrm{MRCA}}$ and $\mathbb{E}T_4^{\mathrm{MRCA}}$, then the right-hand sides of (4) and (7) must agree, while the right-hand sides of (5) and (8) must agree. This implies

$$\frac{f_1(d) + e^{5/d}\mathrm{Ei}(-(6/d))f_2(d)}{\mathrm{Ei}(-(1/d))f_2(d)} = 0, \tag{9}$$

where

$$\begin{aligned}f_1(d) :=\ & 2\mathrm{Ei}\left(-\frac{1}{d}\right)\left\{e^{4/d}\left[\mathrm{Ei}\left(-\frac{3}{d}\right)\right]^2 - 4e^{2/d}\mathrm{Ei}\left(-\frac{3}{d}\right)\mathrm{Ei}\left(-\frac{1}{d}\right)\right. \\ &\left. + \left[\mathrm{Ei}\left(-\frac{1}{d}\right)\right]^2\right\}, \\ f_2(d) :=\ & 3e^{4/d}\left[\mathrm{Ei}\left(-\frac{3}{d}\right)\right]^2 - 2e^{2/d}\mathrm{Ei}\left(-\frac{3}{d}\right)\mathrm{Ei}\left(-\frac{1}{d}\right) \\ &+ 3\left[\mathrm{Ei}\left(-\frac{1}{d}\right)\right]^2.\end{aligned}$$

However, by *Lemma 8* in the *Appendix*, there is no $d \in (0, \infty)$ such that (9) holds.

Case 3. One Λ measure is a β-coalescent and the other is Kingman's coalescent with exponential growth. If these two coalescents produce the same values of $\mathbb{E}T_2^{\mathrm{MRCA}}$, $\mathbb{E}T_3^{\mathrm{MRCA}}$, and $\mathbb{E}T_4^{\mathrm{MRCA}}$, then we must have $t = 3/2 - e^{2/d}\mathrm{Ei}(-3/d)/2\mathrm{Ei}(-1/d)$ in (6), and equating (6) and (8) implies

$$\frac{g_1(d) + 3e^{5/d}\mathrm{Ei}(-(6/d))g_2(d)}{\mathrm{Ei}(-(1/d))g_2(d)} = 0, \tag{10}$$

where

$$\begin{aligned}g_1(d) &:= 2\mathrm{Ei}\left(-\frac{1}{d}\right)\left[-4e^{2/d}\mathrm{Ei}\left(-\frac{3}{d}\right) + \mathrm{Ei}\left(-\frac{1}{d}\right)\right], \\ g_2(d) &:= e^{2/d}\mathrm{Ei}\left(-\frac{3}{d}\right) + \mathrm{Ei}\left(-\frac{1}{d}\right).\end{aligned}$$

However, by *Lemma 9* in the *Appendix*, there is no $d \in (0, \infty)$ such that (10) holds.

Since each of the three cases results in a contradiction, we see that no such Λ measures exist, proving the identifiability claim.

□

Discussion

We have presented an efficient algorithm for computing the SFS for a very general class of coalescents. While Λ- and Ξ-coalescents seem to be primarily used in practice to model the genealogies of marine species (Árnason 2004; Hedgecock and Pudovkin 2011), these coalescents also model a wide range of other phenomena, including continuous strong positive selection (Neher and Hallatschek 2013), recurrent selective sweeps (Durrett and Schweinsberg 2004, 2005), strong bottlenecks (Birkner *et al.* 2009), and many others. Perhaps one of the reasons these coalescents are less widely used than Kingman's coalescent is because efficient inference tools have not yet been developed to the same extent.

Multiple-merger coalescents have also attracted some interest recently in the context of extremely large sample sizes (Bhaskar *et al.* 2014). In such cases the sample size is too large for the assumption of only pairwise mergers of lineages imposed by Kingman's coalescent to be biologically plausible, and indeed using Kingman's coalescent to model such populations causes biases in inference (Bhaskar *et al.* 2014). It should be possible to extend the results presented in this article to discrete-time coalescents, such as the "exact coalescent" (Fu 2006) corresponding to the coalescent arising from the discrete-time Wright–Fisher process, or to any of the discrete-time random-mating models considered by Eldon and Wakeley (2006).

We also presented some encouraging identifiability results. While it is impossible in the general case to infer the inverse intensity function ζ or the measure of a Λ-coalescent from the SFS, for many biologically important cases identifiability does indeed hold. The method we presented for proving that the Λ measure is identifiable for constant ζ is powerful, but straightforward and should make it easy to prove whether the measure is identifiable for other sets of Λ- or Ξ-coalescents. While we considered the identifiability of Λ only for fixed, constant ζ and the identifiability of ζ for fixed Λ or Ξ, it would be interesting to see whether identifiability results can still be obtained for some model spaces while allowing both Λ and ζ to vary. It would also be interesting to extend our identifiability results for the Λ measure to some of the biologically relevant Ξ-coalescents.

Our identifiability results generally assumed access to the expected SFS. In practice, one observes a finite number of sites and so one has only a noisy estimate of the SFS. Our simulation study shows that, given a moderate number of independent trees, the empirical SFS is accurate enough to distinguish $\Lambda(dx)/\eta(t)$ for some simple models. However, the effects of noisy data are still largely unknown, especially in cases where convergence to the expected SFS is not guaranteed. The accuracy of inferring ζ with the empirical SFS has been studied for Kingman's coalescent (Terhorst and Song 2015), and it would be interesting to extend these results to general Λ-coalescents and to the inference of the Λ-measure itself; the results presented here should make such an analysis more tractable.

Acknowledgments

We thank Jere Koskela for helpful discussion on convergence to the expected SFS. This research is supported in part by National Institutes of Health (NIH) grant R01-GM108805, NIH training grant T32-HG000047, and a Packard Fellowship for Science and Engineering.

Literature Cited

Árnason, E., 2004 Mitochondrial cytochrome *B* DNA variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. Genetics 166: 1871–1885.

Berestycki, J., N. Berestycki, and J. Schweinsberg, 2007 Beta-coalescents and continuous stable random trees. Ann. Probab. 35: 1835–1887.

Berestycki, J., N. Berestycki, and V. Limic, 2014 Asymptotic sampling formulae for Λ-coalescents, pp. 715–731 in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Vol. 50. Institut Henri Poincaré. Available at: http://projecteuclid.org/euclid.aihp/1403276996.

Berestycki, N., 2009 Recent progress in coalescent theory. Ensaios Matematicos 16(1): 1–193.

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

Bhaskar, A., A. G. Clark, and Y. S. Song, 2014 Distortion of genealogical properties when the sample is very large. Proc. Natl. Acad. Sci. USA 111(6): 2385–2390.

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 Res. 25(2): 268–279.

Birkner, M., J. Blath, M. Möhle, M. Steinrücken, and J. Tams, 2009 A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. Alea 6: 25–61.

Birkner, M., J. Blath, and B. Eldon, 2013a Statistical properties of the site-frequency spectrum associated with λ-coalescents. Genetics 195: 1037–1053.

Birkner, M., J. Blath, and B. Eldon, 2013b An ancestral recombination graph for diploid populations with skewed offspring distribution. Genetics 193: 255–290.

Blath, J., M. C. Cronjäger, B. Eldon, and M. Hammer, 2015 The site-frequency spectrum associated with Ξ-coalescents. bioRxiv Preprint. Available at: http://dx.doi.org/10.1101/025684.

Bolthausen, E., and A. S. Sznitman, 1998 On Ruelle's probability cascades and an abstract cavity method. Commun. Math. Phys. 197: 247–276.

Coventry, A., L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell *et al.*, 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat. Commun. 1: 131.

Donnelly, P., and T. G. Kurtz, 1999 Particle representations for measure-valued population models. Ann. Probab. 27(1): 166–205.

Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Popul. Biol. 66(2): 129–138.

Durrett, R., and J. Schweinsberg, 2005 A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stoch. Proc. Appl. 115: 1628–1657.

Eldon, B., and J. Wakeley, 2006 Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172: 2621–2633.

Eldon, B., M. Birkner, J. Blath, and F. Freund, 2015 Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? Genetics 199: 841–856.
Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust demographic inference from genomic and SNP data. PLoS Genet. 9(10): e1003905.
Fu, Y.-X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. 48(2): 172–197.
Fu, Y.-X., 2006 Exact coalescent for the Wright-Fisher model. Theor. Popul. Biol. 69(4): 385–394.
Gao, F., and A. Keinan, 2016 Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. Genetics 202: 235–245.
Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth *et al.*; 1000 Genomes Project, 2011 Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA 108(29): 11983–11988.
Griffiths, R., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Commun. Stat. Stoch. Models 14(1–2): 273–295.
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 Genet. 5(10): e1000695.
Hedgecock, D., and A. I. Pudovkin, 2011 Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. Bull. Mar. Sci. 87(4): 971–1002.
Huillet, T. E., 2014 Pareto genealogies arising from a Poisson branching evolution model with selection. J. Math. Biol. 68 (3): 727–761.
Kamm, J. A., J. Terhorst, and Y. S. Song, 2015 Efficient computation of the joint sample frequency spectra for multiple populations. arXiv Preprint. Available at: http://arxiv.org/abs/1503.01133.
Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. 13: 235–248.
Koskela, J., P. A. Jenkins, and D. Spanò, 2015 Bayesian nonparametric inference for Λ-coalescents: consistency and a parametric method. arXiv Preprint. Available at: http://arxiv.org/abs/1512.00982.
Möhle, M., and H. Pitters, 2014 A spectral decomposition for the block counting process of the Bolthausen-Sznitman coalescent. Electron. Commun. Probab. 19(47): 1–11.
Möhle, M., and S. Sagitov, 2001 A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29(4): 1547–1562.
Möhle, M., and S. Sagitov, 2003 Coalescent patterns in diploid exchangeable population models. J. Math. Biol. 47(4): 337–352.
Myers, S., C. Fefferman, and N. Patterson, 2008 Can one learn history from the allelic spectrum? Theor. Popul. Biol. 73(3): 342–348.
Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. Proc. Natl. Acad. Sci. USA 110(2): 437–442.
Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154: 931–942.
Pitman, J., 1999 Coalescents with multiple collisions. Ann. Probab. 27: 1870–1902.
Polanski, A., and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165: 427–436.
Polanski, A., A. Bobrowski, and M. Kimmel, 2003 A note on distributions of times to coalescence, under time-dependent population size. Theor. Popul. Biol. 63(1): 33–40.
Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Probab. 36(4): 1116–1125.
Schweinsberg, J., 2000 Coalescents with simultaneous multiple collisions. Electron. J. Probab. 5: 1–50.
Schweinsberg, J., 2015 Rigorous results for a population model with selection II: genealogy of the population. arXiv Preprint. Available at: http://arxiv.org/abs/1507.00394.
Terhorst, J., and Y. S. Song, 2015 Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. Proc. Natl. Acad. Sci. USA 112(25): 7677–7682.

Communicating editor: N. H. Barton

Appendix

Here we present two lemmas that are used in *Theorem 5*. Proofs are tedious but straightforward.

Lemma 8. *For* $d \in (0, \infty)$,

$$\frac{f_1(d) + e^{5/d}\mathrm{Ei}(-(6/d))f_2(d)}{\mathrm{Ei}(-(1/d))f_2(d)} \neq 0,$$

where

$$f_1(d) := 2\mathrm{Ei}\left(-\frac{1}{d}\right)\left\{e^{4/d}\left[\mathrm{Ei}\left(-\frac{3}{d}\right)\right]^2 - 4e^{2/d}\mathrm{Ei}\left(-\frac{3}{d}\right)\mathrm{Ei}\left(-\frac{1}{d}\right) + \left[\mathrm{Ei}\left(-\frac{1}{d}\right)\right]^2\right\},$$

$$f_2(d) := 3e^{4/d}\left[\mathrm{Ei}\left(-\frac{3}{d}\right)\right]^2 - 2e^{2/d}\mathrm{Ei}\left(-\frac{3}{d}\right)\mathrm{Ei}\left(-\frac{1}{d}\right) + 3\left[\mathrm{Ei}\left(-\frac{1}{d}\right)\right]^2.$$

Lemma 9. *For* $d \in (0, \infty)$,

$$\frac{g_1(d) + 3e^{5/d}\mathrm{Ei}(-(6/d))g_2(d)}{\mathrm{Ei}(-(1/d))g_2(d)} \neq 0,$$

where

$$g_1(d) := 2\mathrm{Ei}\left(-\frac{1}{d}\right)\left[-4e^{2/d}\mathrm{Ei}\left(-\frac{3}{d}\right) + \mathrm{Ei}\left(-\frac{1}{d}\right)\right],$$

$$g_2(d) := e^{2/d}\mathrm{Ei}\left(-\frac{3}{d}\right) + \mathrm{Ei}\left(-\frac{1}{d}\right).$$

In what follows, let $\mathrm{E}_1(x) := \int_x^\infty (e^{-t}/t)dt = -\mathrm{Ei}(-x)$. It is clear that $\mathrm{E}_1(x) > 0$ for all $x > 0$. Additionally,

$$e^{n/d}\mathrm{E}_1\left(\frac{n+1}{d}\right) = \int_{1/d}^\infty \frac{e^{-t}}{t + n/d}dt, \tag{A1}$$

which follows from the definition of E_1 and a change of variables.

Proof of Lemma 8. First, by noting that $f_2(d) = 3[e^{2/d}\mathrm{E}_1(3/d) - \mathrm{E}_1(1/d)]^2 + 4e^{2/d}\mathrm{E}_1(3/d)\mathrm{E}_1(1/d)$, it is easy to see that the denominator is strictly negative for $d \in (0, \infty)$. We now show that the numerator is strictly positive for $d \in (0, \infty)$. First, by rearranging terms we see that

$$\begin{aligned} f_1(d) + e^{5/d}\mathrm{Ei}\left(-\frac{6}{d}\right)f_2(d) = {} & 4\left[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{5/d}\mathrm{E}_1\left(\frac{6}{d}\right)\right]e^{2/d}\mathrm{E}_1\left(\frac{3}{d}\right)\mathrm{E}_1\left(\frac{1}{d}\right) \\ & - \left[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{2/d}\mathrm{E}_1\left(\frac{3}{d}\right)\right]^2\left[2\mathrm{E}_1\left(\frac{1}{d}\right) + 3e^{5/d}\mathrm{E}_1\left(\frac{6}{d}\right)\right]. \end{aligned} \tag{A2}$$

Then, note

$$\mathrm{E}_1\left(\frac{1}{d}\right) - e^{2/d}\mathrm{E}_1\left(\frac{3}{d}\right) = \frac{2}{d}\int_{1/d}^\infty \frac{e^{-t}}{t(t + 2/d)}dt < \frac{4}{d}\int_{1/d}^\infty \frac{e^{-t}}{t(t + 5/d)} = \frac{4}{5}\left[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{5/d}\mathrm{E}_1\left(\frac{6}{d}\right)\right].$$

Applying this inequality to the negative term on the right-hand side of (12), we see

$$
\begin{aligned}
&f_1(d) + e^{5/d}\mathrm{Ei}\left(-\frac{6}{d}\right) f_2(d) \\
&> 4\left[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{5/d}\mathrm{E}_1\left(\frac{6}{d}\right)\right] \\
&\times \left\{ e^{2/d}\mathrm{E}_1\left(\frac{3}{d}\right)\mathrm{E}_1\left(\frac{1}{d}\right) - \left[\frac{4}{5d}\mathrm{E}_1\left(\frac{1}{d}\right) + \frac{6}{5d}\, e^{5/d}\mathrm{E}_1\left(\frac{6}{d}\right)\right]\left(\int_{1/d}^{\infty} \frac{e^{-t}}{t(t+2/d)}\, dt\right)\right\} \\
&> 4\left[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{5/d}\mathrm{E}_1\left(\frac{6}{d}\right)\right]\mathrm{E}_1\left(\frac{1}{d}\right)\left[e^{2/d}\mathrm{E}_1\left(\frac{3}{d}\right) - \frac{4}{15}\mathrm{E}_1\left(\frac{1}{d}\right) - \frac{2}{5}\, e^{5/d}\mathrm{E}_1\left(\frac{6}{d}\right)\right] \\
&= 4\left[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{5/d}\mathrm{E}_1\left(\frac{6}{d}\right)\right]\mathrm{E}_1\left(\frac{1}{d}\right)\left[\int_{1/d}^{\infty} \frac{((1/3)t^2 + (7/3d)t - 8/3d^2)e^{-t}}{t(t+2/d)(t+5/d)}\right],
\end{aligned}
$$

which is >0 for any $d \in (0, \infty)$ since $\mathrm{E}_1(1/d) > e^{5/d}\mathrm{E}_1(6/d)$ and $(1/3)t^2 + (7/3d)t - 8/3d^2 > 0$ for $t > 1/d$.

□

Proof of Lemma 9. The denominator of (10) is equal to $\mathrm{E}_1(1/d)[e^{2/d}\mathrm{E}_1(3/d) + \mathrm{E}_1(1/d)]$, which is strictly positive for $d \in (0, \infty)$, by definition of $\mathrm{E}_1(x)$. Furthermore, the numerator is strictly negative for $d \in (0, \infty)$ by noting the following:

$$
\begin{aligned}
&g_1(d) + 3e^{5/d}\mathrm{Ei}\left(-\frac{6}{d}\right) g_2(d) \\
&= \left(\int_{1/d}^{\infty} \frac{e^{-t}}{t}\, dt\right)\left[\int_{1/d}^{\infty} \frac{2e^{-t}}{t} + \frac{3e^{-t}}{t+5/d} - \frac{8e^{-t}}{t+2/d}\, dt\right] + 3\left(\int_{1/d}^{\infty} \frac{e^{-t}}{t+5/d}\, dt\right)\left(\int_{1/d}^{\infty} \frac{e^{-t}}{t+2/d}\, dt\right) \\
&= \left(\int_{1/d}^{\infty} \frac{e^{-t}}{t}\, dt\right)\left[\int_{1/d}^{\infty} \frac{(-3t^2 - (20/d)t + 20/d^2)e^{-t}}{t(t+2/d)(t+5/d)}\, dt\right] \\
&\quad + 3\left(\int_{1/d}^{\infty} \frac{e^{-t}}{t+5/d}\, dt\right)\left[\left(\int_{1/d}^{\infty} \frac{e^{-t}}{t}\, dt\right) - \left(\int_{1/d}^{\infty} \frac{(2/d)e^{-t}}{t(t+2/d)}\, dt\right)\right] \\
&= \left(\int_{1/d}^{\infty} \frac{e^{-t}}{t}\, dt\right)\left[\int_{1/d}^{\infty} \frac{(-(14/d)t + 20/d^2)e^{-t}}{t(t+2/d)(t+5/d)}\, dt\right] - \frac{6}{d}\left(\int_{1/d}^{\infty} \frac{e^{-t}}{t+5/d}\, dt\right)\left(\int_{1/d}^{\infty} \frac{e^{-t}}{t(t+2/d)}\, dt\right) \\
&\quad < \left(\int_{1/d}^{\infty} \frac{e^{-t}}{t}\, dt\right)\left[\int_{1/d}^{\infty} \frac{(-(14/d)t + 20/d^2)e^{-t}}{t(t+2/d)(t+5/d)}\, dt\right] - \frac{1}{d}\left(\int_{1/d}^{\infty} \frac{e^{-t}}{t}\, dt\right)\left(\int_{1/d}^{\infty} \frac{e^{-t}}{t(t+2/d)}\, dt\right) \\
&= \left(\int_{1/d}^{\infty} \frac{e^{-t}}{t}\, dt\right)\left[\int_{1/d}^{\infty} \frac{(-(15/d)t + 15/d^2)e^{-t}}{t(t+2/d)(t+5/d)}\, dt\right] \\
&\quad < 0.
\end{aligned}
$$

Therefore, (10) holds.

□