The Annals of Statistics
2014, Vol. 42, No. 6, 2469–2493
DOI: 10.1214/14-AOS1264

DESCARTES' RULE OF SIGNS AND THE IDENTIFIABILITY OF POPULATION DEMOGRAPHIC MODELS FROM GENOMIC VARIATION DATA[1]

BY ANAND BHASKAR AND YUN S. SONG

University of California, Berkeley

The sample frequency spectrum (SFS) is a widely-used summary statistic of genomic variation in a sample of homologous DNA sequences. It provides a highly efficient dimensional reduction of large-scale population genomic data and its mathematical dependence on the underlying population demography is well understood, thus enabling the development of efficient inference algorithms. However, it has been recently shown that very different population demographies can actually generate the same SFS for arbitrarily large sample sizes. Although in principle this nonidentifiability issue poses a thorny challenge to statistical inference, the population size functions involved in the counterexamples are arguably not so biologically realistic. Here, we revisit this problem and examine the identifiability of demographic models under the restriction that the population sizes are *piecewise-defined* where each piece belongs to some family of biologically-motivated functions. Under this assumption, we prove that the expected SFS of a sample uniquely determines the underlying demographic model, provided that the sample is sufficiently large. We obtain a general bound on the sample size sufficient for identifiability; the bound depends on the number of pieces in the demographic model and also on the type of population size function in each piece. In the cases of piecewise-constant, piecewise-exponential and piecewise-generalized-exponential models, which are often assumed in population genomic inferences, we provide explicit formulas for the bounds as simple functions of the number of pieces. Lastly, we obtain analogous results for the "folded" SFS, which is often used when there is ambiguity as to which allelic type is ancestral. Our results are proved using a generalization of Descartes' rule of signs for polynomials to the Laplace transform of piecewise continuous functions.

1. Introduction. Given a sample of homologous genomic sequences from a large population, an important inference problem with a wide variety of important applications is to determine the underlying demography of the population. The population demography can be used to calibrate null models of neutral genome

Received November 2013; revised May 2014.

[1]Supported in part by an NIH Grant R01-GM094402, and a Packard Fellowship for Science and Engineering.

MSC2010 subject classifications. Primary 62B10; secondary 92D15.

Key words and phrases. Population genetics, identifiability, population size, coalescent theory, frequency spectrum.

evolution in order to find regions under selection [2, 25, 45]; to stratify samples in genome-wide association studies [3, 28, 33, 37]; to date historical population splits, migrations, admixture and introgression events [10, 18, 24, 26, 40, 43]; and so on. Recently, several large-sample genome- and exome-sequencing datasets have become available [1, 4, 6, 31, 44], shedding new light on patterns of genetic variation that were not previously observable in smaller datasets. Such large-sample studies offer an exciting opportunity to infer demography in unprecedented detail.

One widely-used measure of genetic variation in a set of homologous genome sequences is the sample frequency spectrum (SFS). For a sample of size n, the SFS counts the proportion of dimorphic (i.e., with exactly two distinct observed alleles) sites as a function of the frequency ($\frac{b}{n}$, where $1 \leq b \leq n-1$) of the mutant allele in the sample. The SFS is useful for several reasons. First, the SFS is a succinct summary of a large sample of genomic sequences, where the information in n sequences of arbitrary length can be summarized by just $n-1$ numbers. This makes the SFS both mathematically and algorithmically tractable. In particular, since the SFS ignores linkage information between sites, one can avoid challenging mathematical and computational issues associated with rigorously modeling genetic recombination. Furthermore, the statistical properties of the SFS and their dependence on the population demographic history are well understood under the coalescent and the diffusion models of neutral evolution [7, 11, 12, 19, 35, 46]. This dependence of the SFS on demography, along with the assumption of free recombination between sites, has been exploited in several efficient methods for inferring historical population demography [5, 13, 27, 29]. Second, the SFS can effectively capture the impact of recent demography on genetic variation. Recent large-sample studies [4, 6, 31, 44] have consistently shown that there is an excess of rare polymorphisms compared to the predictions of previously inferred demographic models, which might be explained by recent rapid population expansion [16]. Because the leading entries of the SFS count the rare variants in the sample, one might be able to use this information to infer demographic events in the recent past at a much finer resolution than possible using smaller samples. Third, the SFS also provides a simple way of visualizing the goodness of fit of a demographic model to data, since one can easily compare the SFS observed in the data with the SFS predicted by the fitted demographic model.

While the SFS has algorithmic advantages for demographic inference, it is believed to suffer from a statistical shortcoming. Specifically, Myers, Fefferman and Patterson [30] recently showed that even with perfect knowledge of the *population* frequency spectrum [i.e., the proportion of polymorphic sites with population-wide allele frequency in $(x, x+dx)$ for all $x \in (0,1)$], the historical population size function $\eta(t)$ as a function of time is not identifiable. Using Müntz–Szász theory, they showed that for any population size function $\eta(t)$, one can construct arbitrarily many smooth functions $F(t)$ such that both $\eta(t)$ and $\eta(t)+\alpha F(t)$ generate the same population frequency spectrum for suitably chosen values of α. They also

constructed explicit examples of such functions $\eta(t)$ and $F(t)$. While this nonidentifiability could pose serious challenges to demographic inference from frequency spectrum data, the population size functions involved in their example are arguably unrealistic for biological populations. In particular, their explicit example involves a population size function which oscillates at an increasingly higher frequency as the time parameter approaches the present. Real biological population sizes can be expected to vary over time in a mathematically more well-behaved fashion. In particular, populations can be expected to evolve in discrete units of time, which, when approximated by a continuous-time model, restricts the frequency of oscillations in the population size function to be less than the number of generations of reproduction per unit time. Furthermore, since a population size model being inferred must have a finite representation for obvious algorithmic reasons, most previous demographic inference analyses have focused on inferring population size models that are piecewise-defined over a restricted class of functions, such as piecewise-constant and piecewise-exponential models [10, 17, 24, 26, 31, 41, 44]. Motivated by the large number of rare variants observed in several large-sample sequencing studies, recent works [38, 39] have also focused on more general population growth models which allow for the population to grow at a faster than exponential rate. Each piece in such piecewise models has two parameters that control the rate and acceleration of population growth. Since these models contain the family of piecewise-constant and piecewise-exponential population size functions, we refer to them as piecewise-generalized-exponential models in the remainder of this paper.

In this paper, we revisit the question of demographic model identifiability under the assumption that the population size is a *piecewise-defined* function of time where each piece comes from a family of biologically-motivated functions, such as the family of constant or exponential functions. We also re-examine the assumption that one has access to the population-wide patterns of polymorphism. In real applications, we do not expect to know the allele frequency spectrum for an entire population but rather only the SFS for a randomly drawn finite sample of individuals. Here, we investigate whether one can learn piecewise-defined population size functions given perfect knowledge of the expected SFS for a sufficiently large sample of size n. Unlike in the case of arbitrary continuous population size functions considered by Myers, Fefferman and Patterson, the answer to this question is affirmative. More precisely, we obtain bounds on the sample size n that are sufficient to distinguish population size functions among piecewise demographic models with K pieces, where each piece comes from some family of functions (see Theorems 6 and 11). Our bound on the sample size can be expressed as an affine function of the number K of pieces, where the slope of the function is a measure of the complexity of the family to which each piece belongs. In the cases of piecewise-constant, piecewise-exponential and piecewise-generalized-exponential models, which are often assumed in population genetic analyses, the slope of this affine function can be calculated explicitly, as shown in Corollaries 7–9. We also

obtain analogous results for the "folded" SFS (see Theorem 12), a variant of the SFS which circumvents the ambiguity in the identity of the ancestral allele type by grouping the polymorphic sites in a sample according to the sample minor allele frequency.

There are two main technical elements underlying our proofs of the identifiability results mentioned above. The first step is to show that the expected SFS of a sample of size n is in bijection with the Laplace transform of a time-rescaled version of the population size function evaluated at a particular sequence of $n-1$ points. This reduces the problem of identifiability from the SFS to that of identifiability from the values of the Laplace transform at a fixed set of points. The second step relies on a generalization of Descartes' rule of signs for polynomials to the Laplace transform of general piecewise-continuous functions. This technique yields an upper bound on the number of roots of the Laplace transform of a function by the number of sign changes of the function. We think that this proof technique based on sign changes might be of independent interest for proving statistical identifiability results in other settings. We also provide an alternate proof of identifiability for piecewise-constant population models, where the aforementioned second step is replaced by a linear algebraic argument that has a constructive flavor. We include this alternate proof in the hope that it could be used to develop an algebraic inference algorithm for piecewise-constant models.

The remainder of this paper is organized as follows. In Section 2, we introduce the model and notation, and describe our main results. We also discuss the counterexample of Myers, Fefferman and Patterson in light of our findings. The proofs of our results are provided in Section 3, and we conclude with a discussion in Section 4.

2. Main results. Here, we summarize our identifiability results. All proofs are deferred to Section 3.

2.1. *Model and notation.* We consider a population evolving according to Kingman's coalescent [21–23] with the infinite-sites model of mutation [20] and selective neutrality. Under this model, the genome is assumed to be infinite and every mutation occurs at a different site in the genome that has never experienced a mutation before. This model is applicable in the regime where the mutation rate is very low, and hence the probability of multiple mutations at a given site is vanishingly small. Any polymorphic site in a sample of sequences is dimorphic under this model. The population size is assumed to change deterministically with time and is described by a function $\eta : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$, such that the instantaneous coalescence rate between any pair of lineages at time t is $1/\eta(t)$.

Let $T_{n,k}^{(\eta)}$ denote the time (in coalescent units) while there are k ancestral lineages for a sample of size n obtained at time 0. Defining $R_\eta(t)$ as

$$R_\eta(t) := \int_0^t \frac{1}{\eta(x)}\, dx,$$

the expected time $\mathbb{E}[T_{m,m}^{(\eta)}]$ to the first coalescence event for a sample of size m is given by

$$\mathbb{E}\big[T_{m,m}^{(\eta)}\big] = \int_0^\infty t \frac{\binom{m}{2}}{\eta(t)} \exp\Big[-\binom{m}{2} R_\eta(t)\Big] dt. \tag{1}$$

Following the notation of Myers, Fefferman and Patterson, define a time-rescaled version $\widetilde{\eta}$ of the population size function η as

$$\widetilde{\eta}(\tau) = \eta\big(R_\eta^{-1}(\tau)\big), \tag{2}$$

where $\tau \in \mathbb{R}_{\geq 0}$. The function $\widetilde{\eta}(\tau)$ reparameterizes the population size as a function of the cumulative rate of coalescence $\tau = R_\eta(t)$. For a given population size function $\widetilde{\eta}$ parameterized by the total coalescence rate τ, there corresponds a unique population size function η parameterized by time t. Specifically, $\eta(t) = \widetilde{\eta}(S_{\widetilde{\eta}}^{-1}(t))$, for all $t \in \mathbb{R}_{\geq 0}$, where $S_{\widetilde{\eta}}(t)$ is an invertible function given by

$$S_{\widetilde{\eta}}(t) = \int_0^t \widetilde{\eta}(x)\, dx.$$

Applying integration by parts to (1) and using the condition that $\mathbb{E}[T_{m,m}^{(\eta)}] < \infty$, we have

$$\mathbb{E}\big[T_{m,m}^{(\eta)}\big] = \int_0^\infty \exp\Big[-\binom{m}{2} R_\eta(t)\Big] dt. \tag{3}$$

Furthermore, since R_η is monotonically increasing and continuous from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$, it is a bijection over $\mathbb{R}_{\geq 0}$. For notational convenience, for any interval $I \subseteq \mathbb{R}_{\geq 0}$, we define $R_\eta(I)$ to be the interval

$$R_\eta(I) = \big\{R_\eta(x) | x \in I\big\}.$$

By making the substitution $\tau = R_\eta(t)$ in (3) and using (2), we have the following expression for $\mathbb{E}[T_{m,m}^{(\eta)}]$:

$$\mathbb{E}\big[T_{m,m}^{(\eta)}\big] = \int_0^\infty \widetilde{\eta}(\tau) \exp\Big[-\binom{m}{2} \tau\Big] d\tau. \tag{4}$$

Equation (4) states that the time to the first coalescence event for a sample of size m is given by the Laplace transform of the time-rescaled population size function $\widetilde{\eta}$ evaluated at the point $\binom{m}{2}$. For a sample of size n, let $\xi_{n,b}$ denote the probability that a dimorphic site has b mutant alleles and $n - b$ ancestral alleles. We refer to $(\xi_{n,1}, \ldots, \xi_{n,n-1})$ as the expected sample frequency spectrum (SFS).

2.2. *Determining the expected times to the first coalescence from the SFS.* The following lemma shows that the expected SFS for a sample of size n tightly constrains the expected time to the first coalescence event for all sample sizes $2, \ldots, n$:

LEMMA 1. *Under an arbitrary variable population size model* $\{\eta(t), t \geq 0\}$, *suppose* $\xi_{n,1}, \ldots, \xi_{n,n-1}$ *are known and define* $c_m := \mathbb{E}[T^{(\eta)}_{m,m}]$ *for* $2 \leq m \leq n$. *Then, up to a common positive multiplicative constant, the quantities* $c_2, \ldots, c_n$ *can be determined uniquely from* $\xi_{n,1}, \ldots, \xi_{n,n-1}$.

This implies that the problem of identifying the population size function $\eta(t)$ from $\xi_{n,1}, \ldots, \xi_{n,n-1}$ can be reduced, up to a multiplicative constant, to the problem of identifying $\eta(t)$ from $c_2, \ldots, c_n$.

2.3. *Piecewise population size models and sign change complexities*. To state our main result in full generality, we first need a few definitions.

DEFINITION 1 ($\mathcal{F}$, family of continuous population size functions). A family $\mathcal{F}$ of continuous population size functions is a set of positive continuous functions $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$ of a particular type parameterized by a collection of variables.

We use $\mathcal{F}_c$ to denote the family of constant population size functions; that is, functions of the form $f(t) = \nu$ for all t, where $\nu \in \mathbb{R}_+$ is the only parameter of the family. Further, we use $\mathcal{F}_e$ to denote the family of exponential population size functions of the form $f(t) = \nu \exp(\beta t)$, where $\nu \in \mathbb{R}_+$ and $\beta \in \mathbb{R}$ are the parameters of the family. In human genetics, there has been recent interest [38, 39] in modeling superexponential growth in the effective population size via models that generalize exponential growth by incorporating an additional acceleration parameter γ. Such population size functions f satisfy the differential equation $df/dt = \beta f(t)^\gamma$ with initial condition $f(0) = \nu$, where $\beta \in \mathbb{R}$, $\gamma \in \mathbb{R}_{\geq 0}$, and $\nu \in \mathbb{R}_+$. When $0 \leq \gamma < 1$ (resp., $\gamma > 1$), this represents superexponential (resp., subexponential) population growth/decline, while $\gamma = 1$ corresponds to exponential population growth/decline. We let $\mathcal{F}_g$ denote the family of such generalized-exponential population size functions.

DEFINITION 2 [$\mathcal{M}_K(\mathcal{F})$, piecewise models over $\mathcal{F}$ with at most K pieces]. Given a family $\mathcal{F}$ of continuous population size functions, a population size function $\eta(t)$ defined over $\mathbb{R}_{\geq 0}$ is said to be piecewise over $\mathcal{F}$ with at most K pieces if there exists an integer p, where $1 \leq p \leq K - 1$, and a sequence of p time points $0 < t_1 < \cdots < t_p < \infty$ such that for each $1 \leq i \leq p + 1$, there exists a positive continuous function $f_i \in \mathcal{F}$ such that $\eta(t) = f_i(t - t_{i-1})$ for all $t \in [t_{i-1}, t_i)$. For convention, we define $t_0 = 0$ and $t_{p+1} = \infty$. Note that η may not be continuous at the change points $t_1, \ldots, t_p$. We use $\mathcal{M}_K(\mathcal{F})$ to denote the space of such piecewise population size models with at most K pieces, each of which belongs to function family $\mathcal{F}$. Illustrated in Figure 1 is an example of piecewise-exponential population size function $\eta \in \mathcal{M}_K(\mathcal{F})$ where $K \geq 5$ and $\mathcal{F} = \mathcal{F}_e$.

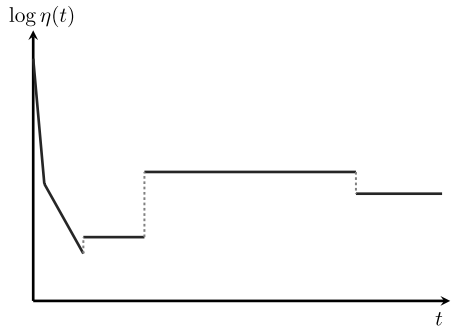

FIG. 1. *A piecewise-exponential population size function* $\eta \in \mathcal{M}_K(\mathcal{F}_e)$, *where* $K \geq 5$. *Note that the y-axis is in a log scale. This piecewise-exponential function depicts the historical population size changes of a European population that was estimated from the SFS of a sample of* 1351 (*diploid*) *individuals of European ancestry* [44].

DEFINITION 3 [$\sigma(f)$, number of sign changes of a function]. For a function g (not necessarily continuous) defined over some interval (a, b), we say that $t \in (a, b)$ is a sign change point of g if there exist some $\varepsilon > 0$, $t' \geq t$, and an interval $(t', t' + \varepsilon) \subseteq (a, b)$ such that:

1. $(t - \varepsilon, t) \subseteq (a, b)$,
2. $g(z) = 0$ for $z \in (t, t')$,
3. $g(x)g(y) < 0$ for all $x \in (t - \varepsilon, t)$ and $y \in (t', t' + \varepsilon)$.

We define the number $\sigma(g)$ of sign changes of g as the number of such sign change points in its domain (a, b). See Figure 2 for an illustration.

Note that the above definition of the number of sign changes counts the number of times the function g changes value from positive to negative (and vice versa) while ignoring intervals where it is identically zero. While the above definition is not restricted to piecewise continuous functions, we will restrict our attention to such functions for the remainder of this paper.

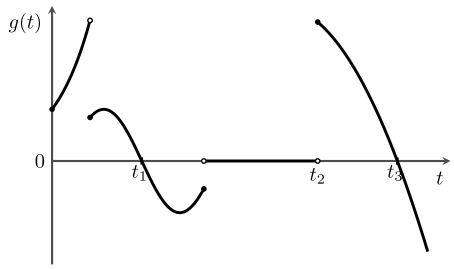

FIG. 2. *Illustration of the sign changes of a function. For the domain shown,* $\sigma(g) = 3$ *and the sign change points of g are denoted* t_1, t_2, *and* t_3.

DEFINITION 4 [$\mathscr{S}(\mathcal{F})$ and $\mathscr{S}(\mathcal{M}_K(\mathcal{F}))$, sign change complexities]. For a family $\mathcal{F}$ of continuous population size functions, we define the sign change complexity $\mathscr{S}(\mathcal{F})$ as

$$\begin{aligned}\mathscr{S}(\mathcal{F}) &= \sup_{\substack{f_1,f_2\in\mathcal{F},\\ a_1,a_2\in\mathbb{R}_{\geq 0}}} \left\{ \sigma(g) \,\middle|\, \begin{array}{l} g(\tau) := \widetilde{f}_1(\tau - a_1) - \widetilde{f}_2(\tau - a_2) \text{ with domain} \\ \mathrm{Dom}(g) = \left\{ \tau \in \mathbb{R}_{\geq 0} \,\middle|\, \begin{array}{l} \tau - a_1 \in \mathrm{Dom}(\widetilde{f}_1), \\ \tau - a_2 \in \mathrm{Dom}(\widetilde{f}_2) \end{array} \right\} \end{array} \right\} \\ &= \sup_{\substack{f_1,f_2\in\mathcal{F},\\ a\in\mathbb{R}_{\geq 0}}} \left\{ \sigma(g) \,\middle|\, \begin{array}{l} g(\tau) := \widetilde{f}_1(\tau) - \widetilde{f}_2(\tau - a) \text{ with domain} \\ \mathrm{Dom}(g) = \left\{ \tau \in \mathbb{R}_{\geq 0} \,\middle|\, \begin{array}{l} \tau \in \mathrm{Dom}(\widetilde{f}_1), \\ \tau - a \in \mathrm{Dom}(\widetilde{f}_2) \end{array} \right\} \end{array} \right\}, \end{aligned} \tag{5}$$

where $\widetilde{f}_j$ are the time-rescaled versions of f_j as defined in (2), and $\mathrm{Dom}(\widetilde{f}_j) = R_{f_j}(\mathbb{R}_{\geq 0})$ is the domain of $\widetilde{f}_j$. Similarly, for the space $\mathcal{M}_K(\mathcal{F})$ of piecewise population size models with at most K pieces over some function family $\mathcal{F}$, we define the sign change complexity $\mathscr{S}(\mathcal{M}_K(\mathcal{F}))$ as

$$\mathscr{S}\big(\mathcal{M}_K(\mathcal{F})\big) = \sup_{\eta_1,\eta_2\in\mathcal{M}_K(\mathcal{F})} \big\{\sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2)\big\},$$

where, again, $\widetilde{\eta}_j$ are related to η_j as given in (2).

The following lemma gives a bound on the sign change complexity of a model with at most K pieces in terms of the underlying family of population size functions for each piece.

LEMMA 2. *The sign change complexity of the space $\mathcal{M}_K(\mathcal{F})$ of piecewise models with at most K pieces in a function family $\mathcal{F}$ is bounded by the sign change complexity of $\mathcal{F}$ as*

$$\mathscr{S}\big(\mathcal{M}_K(\mathcal{F})\big) \leq (2K-2) + (2K-1)\mathscr{S}(\mathcal{F}).$$

Note that the bound in Lemma 2 is tight for the family $\mathcal{F}_c$ of constant population sizes, for which $\mathscr{S}(\mathcal{F}_c) = 0$ and $\mathscr{S}(\mathcal{M}_K(\mathcal{F}_c)) = 2K - 2$.

2.4. *Identifiability results.* Our main results on identifiability will be proved using a generalization of Descartes' rule of signs for polynomials.

THEOREM 3 (Descartes' rule of signs for polynomials). *Consider a degree-n polynomial $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ with real-valued coefficients a_i. The number of positive real roots (counted with multiplicity) of p is at most the number of sign changes between consecutive nonzero terms in the sequence $a_0, a_1, \ldots, a_n$.*

The following theorem generalizes the above classic result to relate the number of sign changes of a piecewise-continuous function f to the number of roots of its Laplace transform.

THEOREM 4 (Generalized Descartes' rule of signs). *Let $f:\mathbb{R}_{\geq 0}\to\mathbb{R}$ be a piecewise-continuous function which is not identically zero and with a finite number $\sigma(f)$ of sign changes. Then the function $G(x)$ defined by*

$$G(x)=\int_0^\infty f(t)e^{-tx}\,dt \tag{6}$$

has at most $\sigma(f)$ roots in $\mathbb{R}$ (counted with multiplicity).

The statement of Theorem 4 and the proof provided in Section 3 are adapted from Jameson [15], Lemma 4.5, for our setting. Using Theorem 4, we prove in Section 3 the following identifiability theorem for population size function families with finite sign change complexity.

THEOREM 5. *For a sample of size n, let $\mathbf{c}=(c_2,\dots,c_n)$, where $c_m=\mathbb{E}[T^{(\eta)}_{m,m}]$, for $2\leq m\leq n$, defined in (3). If $\mathscr{S}(\mathcal{F})<\infty$ and $n\geq\mathscr{S}(\mathcal{F})+2$, then no two distinct models $\eta_1,\eta_2\in\mathcal{F}$ can produce the same $(c_2,\dots,c_n)$. In other words, for $n\geq\mathscr{S}(\mathcal{F})+2$, the map $\mathbf{c}:\mathcal{F}\to\mathbb{R}_+^{n-1}$ is injective.*

Note that the sample size bound in Theorem 5 applies to an arbitrary function family $\mathcal{F}$ which need not have any special structure. Using Lemma 2 for bounding the sign change complexity of piecewise-defined function families $\mathcal{M}_K(\mathcal{F})$ in terms of the sign change complexity of the underlying function family $\mathcal{F}$, we immediately obtain the following theorem.

THEOREM 6. *For a sample of size n, let $\mathbf{c}=(c_2,\dots,c_n)$, where $c_m=\mathbb{E}[T^{(\eta)}_{m,m}]$, for $2\leq m\leq n$, defined in (3). If $\mathscr{S}(\mathcal{F})<\infty$ and $n\geq 2K+(2K-1)\mathscr{S}(\mathcal{F})$, then the map $\mathbf{c}:\mathcal{M}_K(\mathcal{F})\to\mathbb{R}_+^{n-1}$ is injective.*

Using Theorem 6, it is simple to derive identifiability results for piecewise-defined population size models over several function families $\mathcal{F}$ that are of biological interest. In particular, we have the following result for the case of piecewise-constant models.

COROLLARY 7 [Identifiability of piecewise-constant population size models in $\mathcal{M}_K(\mathcal{F}_c)$]. *The map $\mathbf{c}:\mathcal{M}_K(\mathcal{F}_c)\to\mathbb{R}_+^{n-1}$ is injective if the sample size $n\geq 2K$.*

The bound in Corollary 7 on the sample size sufficient for identifying piecewise-constant population models is actually tight, since $\mathcal{M}_K(\mathcal{F}_c)$ has $2K-1$ parameters in $\mathbb{R}_+$ and there is no continuous injective function from $\mathbb{R}_+^{2K-1}$ to $\mathbb{R}^{n-1}$ if $n<2K$. (This fact can be proved in multiple ways, such as by the Borsuk–Ulam theorem or the Constant Rank theorem.) An alternate proof of Corollary 7 that does not rely on Theorem 6 is also provided in Section 3. This alternate proof

is based on an argument from linear algebra, and it might be possible to adapt this approach to develop an algebraic algorithm for inferring the parameters of a piecewise-constant population function from the set of expected first coalescence times c_m.

Another class of models often assumed in population genetic analyses are piecewise-exponential functions, for which we have the following result.

COROLLARY 8 [Identifiability of piecewise-exponential population size models in $\mathcal{M}_K(\mathcal{F}_e)$]. *The map* $\mathbf{c}:\mathcal{M}_K(\mathcal{F}_e)\to\mathbb{R}_+^{n-1}$ *is injective if the sample size* $n\geq 4K-1$.

For the generalized-exponential growth models considered by Reppell, Boehnke and Zöllner [39], we have the following result.

COROLLARY 9 [Identifiability of piecewise-generalized-exponential population size models in $\mathcal{M}_K(\mathcal{F}_g)$]. *The map* $\mathbf{c}:\mathcal{M}_K(\mathcal{F}_g)\to\mathbb{R}_+^{n-1}$ *is injective if the sample size* $n\geq 6K-2$.

For the identifiability of piecewise population size models from the SFS data, we first note the following lemma.

LEMMA 10. *Consider a piecewise population size function* $\eta\in\mathcal{M}_K(\mathcal{F})$. *Consider a sample of size* $n\geq 2K+(2K-1)\mathscr{S}(\mathcal{F})$ *and suppose the function* η *produces* $\mathbb{E}[T_{m,m}^{(\eta)}]=c_m$ *for* $2\leq m\leq n$. *Then, for every fixed* $\kappa\in\mathbb{R}_+$, *there exists a unique piecewise population size function* $\zeta\in\mathcal{M}_K(\mathcal{F})$ *with* $\mathbb{E}[T_{m,m}^{(\zeta)}]=\kappa c_m$ *for* $2\leq m\leq n$. *Furthermore, this population size function* ζ *is given by* $\zeta(t)=\kappa\eta(t/\kappa)$.

Given two models $\eta,\zeta\in\mathcal{M}_K$, we say that η and ζ are equivalent, and write $\eta\sim\zeta$, if they are related by a rescaling of change points and population sizes as described in Lemma 10. Let $[\eta]$ denote the equivalence class of population size functions that contain η, and let $\mathcal{M}_K(\mathcal{F})/\sim=\{[\eta]|\eta\in\mathcal{M}_K(\mathcal{F})\}$ be the set of equivalence classes for the equivalence relation $\sim$. Then, combining Lemma 1, Theorem 6 and Lemma 10, we obtain the following theorem.

THEOREM 11. *If* $\mathscr{S}(\mathcal{F})<\infty$ *and* $n\geq 2K+(2K-1)\mathscr{S}(\mathcal{F})$, *then, for each expected SFS* $(\xi_{n,1},\ldots,\xi_{n,n-1})$, *there exists a unique equivalence class* $[\eta]$ *of models in* $\mathcal{M}_K(\mathcal{F})/\sim$ *consistent with* $(\xi_{n,1},\ldots,\xi_{n,n-1})$.

2.5. *Extension to the folded frequency spectrum.* To generate the SFS from genomic sequence data, one needs to know the identities of the ancestral and mutant alleles at each site. To avoid this problem, a commonly employed strategy

in population genetic inference involves "folding" the SFS. More precisely, for a sample of size n, the ith entry of the folded SFS $\boldsymbol{\chi} = (\chi_{n,1}, \ldots, \chi_{\lfloor n/2 \rfloor})$ is defined by

$$\chi_{n,i} = \frac{\xi_{n,i} + \xi_{n,n-i}}{1 + \delta_{i,n-i}},$$

where $1 \le i \le \lfloor n/2 \rfloor$. In particular, $\chi_{n,i}$ is the proportion of polymorphic sites that have i copies of the minor allele. For any sample size n, since $\boldsymbol{\chi}$ is a vector of approximately half the dimension as $\boldsymbol{\xi}$, we might expect to require roughly twice as many samples to recover the demographic model from $\boldsymbol{\chi}$ compared to $\boldsymbol{\xi}$. This is indeed the case. Given the folded SFS $\boldsymbol{\chi}$, the following theorem establishes a sufficiency condition on the sample size for identifying demographic models in $\mathcal{M}_K(\mathcal{F})$.

THEOREM 12. *If $\mathscr{S}(\mathcal{F}) < \infty$ and $n \ge 2(2K-1)(1+\mathscr{S}(\mathcal{F}))$, then, for each expected folded SFS $\boldsymbol{\chi} = (\chi_{n,1}, \ldots, \chi_{n,\lfloor n/2 \rfloor})$, there exists a unique equivalence class $[\eta]$ of models in $\mathcal{M}_K(\mathcal{F})/\sim$ consistent with $\boldsymbol{\chi}$.*

2.6. *The counterexample of Myers, Fefferman and Patterson.* Myers, Fefferman and Patterson [30] provided an explicit counterexample to the identifiability of population size models from the allelic frequency spectrum. In our notation, they provided two time-rescaled population size functions $\widetilde{\eta}_1$ and $\widetilde{\eta}_2$ given by

$$\begin{aligned}\widetilde{\eta}_1(\tau) &= N,\\ \widetilde{\eta}_2(\tau) &= N\bigl(1 - 9F(\tau)\bigr),\end{aligned}$$

where N is an arbitrary positive constant, and the function F is given by the convolution

$$F(\tau) = \int_0^\tau f_0(\tau - u) f_1(u)\, du,$$

where f_0 and f_1 are given by

$$\begin{aligned}f_0(\tau) &= \exp\bigl(-1/\tau^2\bigr),\\ f_1(\tau) &= \frac{\cos(\pi^2/\tau)\exp(-\tau/8)}{\sqrt{\tau}}.\end{aligned}$$

Both functions f_1 and F have increasingly frequent oscillations as $\tau \downarrow 0$ so that $\sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2) = \sigma(F) = \infty$. This is why Theorem 5 does not apply to this example. Indeed, by an argument using the Laplace transforms of f_1 and F, Myers, Fefferman and Patterson showed that the function $G(x)$ defined in (6) in terms of F has roots at $-\binom{m}{2}$ for each $m \ge 2$.

3. Proofs. We now provide proofs of the results presented earlier.

PROOF OF LEMMA 1. In the coalescent for a sample of size n, let $\gamma_{n,b}$ denote the total expected branch length subtending b leaves, for $1 \le b \le n-1$. Then $\xi_{n,b} = \gamma_{n,b}/\sum_{k=1}^{n-1} \gamma_{n,k}$, which implies that there exists a positive constant κ such that $\gamma_{n,b} = \kappa \xi_{n,b}$ for all $1 \le b \le n-1$. We now prove that $c_2, \ldots, c_n$ can be determined uniquely from $\gamma_{n,1}, \ldots, \gamma_{n,n-1}$.

Let $\phi_{n,k} = \mathbb{E}[T_{n,k}^{(\eta)}]$. Then, by a result of Griffiths and Tavaré [12],

$$\gamma_{n,b} = \sum_{k=2}^{n-b+1} k \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \phi_{n,k}, \tag{7}$$

for $1 \le b \le n-1$. The system of equations (7) can be rewritten succinctly as a linear system

$$\boldsymbol{\gamma} = \mathbf{M}\boldsymbol{\phi},$$

where $\boldsymbol{\gamma} = (\gamma_{n,1}, \ldots, \gamma_{n,n-1})$, $\phi = (\phi_{n,2}, \ldots, \phi_{n,n})$, and $\mathbf{M} = (m_{bk})$ with $m_{bk} = k\binom{n-b-1}{k-2}/\binom{n-1}{k-1}$, for $1 \le b \le n-1$ and $2 \le k \le n$. The matrix $\mathbf{M}$ is upper-left triangular since $\binom{n-b-1}{k-2} = 0$ if $k > n-b+1$, and the anti-diagonal entries are $\frac{k}{\binom{n-1}{k-1}} > 0$. Hence, $\det(\mathbf{M}) \neq 0$ and $\mathbf{M}$ is therefore invertible. Thus, given $\boldsymbol{\gamma}$, we can determine $\boldsymbol{\phi}$ uniquely as $\mathbf{M}^{-1}\boldsymbol{\gamma}$.

Let $\psi_{n,k} = \sum_{j=k}^{n} \mathbb{E}[T_{n,j}^{(\eta)}]$. Then, defining $\psi_{n,n+1} := 0$, observe that $\psi_{n,k} = \phi_{n,k} + \psi_{n,k+1}$ for $2 \le k \le n$. This implies that $\psi_{n,2}, \ldots, \psi_{n,n}$ can be determined uniquely from $\phi_{n,2}, \ldots, \phi_{n,n}$. Polanski, Bobrowski and Kimmel [35] showed that $\psi_{n,k}$ can be written as

$$\psi_{n,k} = \sum_{m=k}^{n} a_{km} c_m, \tag{8}$$

where a_{km}, for $k \le m \le n$, are given by

$$a_{km} = \frac{\prod_{l=k, l\neq m}^{n} \binom{l}{2}}{\prod_{l=k, l\neq m}^{n} [\binom{l}{2} - \binom{m}{2}]},$$

and $c_m = \mathbb{E}[T_{m,m}^{(\eta)}]$, shown in (3). Again, the system of equations (8) can be written as a triangular linear system

$$\boldsymbol{\psi} = \mathbf{A}\mathbf{c},$$

where $\boldsymbol{\psi} = (\psi_{n,2}, \ldots, \psi_{n,n})$, $\mathbf{c} = (c_2, \ldots, c_n)$, and $\mathbf{A} = (a_{km})$, for $2 \le k, m \le n$. Note that $\mathbf{A}$ is an upper triangular matrix since $a_{km} := 0$ if $m < k$. Since $\mathbf{A}$ has nonzero entries on its diagonal, $\mathbf{A}^{-1}$ exists, and $\mathbf{c}$ can be determined uniquely as $\mathbf{A}^{-1}\boldsymbol{\psi}$. $\square$

PROOF OF LEMMA 2. Given a pair of piecewise population size functions $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$, let $\widetilde{\eta}_1$ and $\widetilde{\eta}_2$ be their respective time-rescaled versions, defined by (2). Let $0 < t_1^{(1)} < \cdots < t_{p_1}^{(1)} < \infty$, where $0 \le p_1 \le K-1$ (resp., $0 < t_1^{(2)} < \cdots < t_{p_2}^{(2)} < \infty$, where $0 \le p_2 \le K-1$) be the change points of the pieces of η_1 (resp., η_2). We define $t_0^{(1)} = t_0^{(2)} = 0$ and $t_{p_1+1}^{(1)} = t_{p_2+1}^{(2)} = \infty$. The change points of $\widetilde{\eta}_1$ are given by $R_{\eta_1}(t_i^{(1)})$, where $1 \le i \le p_1$, while the change points of $\widetilde{\eta}_2$ are given by $R_{\eta_2}(t_i^{(2)})$, where $1 \le i \le p_2$. Let $0 < \tau_1 < \cdots < \tau_p < \infty$ be the union of the change points of $\widetilde{\eta}_1$ and $\widetilde{\eta}_2$, where $0 \le p \le p_1 + p_2$. For convention, let $\tau_0 = 0$ and $\tau_{p+1} = \infty$.

Consider the piece (τ_i, τ_{i+1}) for $0 \le i \le p$. Let $I_1 = (t_k^{(1)}, t_{k+1}^{(1)})$, where $0 \le k \le p_1$, and $I_2 = (t_l^{(2)}, t_{l+1}^{(2)})$, where $0 \le l \le p_2$, be the pieces of the original population size functions η_1 and η_2, respectively, such that $(\tau_i, \tau_{i+1}) \subseteq R_{\eta_1}(I_1)$ and $(\tau_i, \tau_{i+1}) \subseteq R_{\eta_2}(I_2)$. Since $\eta_1 \in \mathcal{M}_K(\mathcal{F})$, there exists a function $f_1 \in \mathcal{F}$ such that $\eta_1(t) = f_1(t - t_k^{(1)})$ for all $t \in I_1$. Then, for all $\tau \in R_{\eta_1}(I_1)$,

$$\begin{aligned} \widetilde{\eta}_1(\tau) &= \eta_1\big(R_{\eta_1}^{-1}(\tau)\big) = f_1\big(R_{\eta_1}^{-1}(\tau) - t_k^{(1)}\big) \\ &= \widetilde{f}_1\big(R_{f_1}\big(R_{\eta_1}^{-1}(\tau) - t_k^{(1)}\big)\big) \\ &= \widetilde{f}_1\big(\tau - R_{\eta_1}\big(t_k^{(1)}\big)\big). \end{aligned} \tag{9}$$

Similarly, there exists some function $f_2 \in \mathcal{F}$ such that, for all $\tau \in R_{\eta_2}(I_2)$,

$$\widetilde{\eta}_2(\tau) = \widetilde{f}_2\big(\tau - R_{\eta_2}\big(t_l^{(2)}\big)\big). \tag{10}$$

Using (9) and (10), we see that the number of sign change points of $\widetilde{\eta}_1 - \widetilde{\eta}_2$ in the piece (τ_i, τ_{i+1}) is at most the number of sign change points of $\widetilde{f}_1(\tau - R_{\eta_1}(t_k^{(1)})) - \widetilde{f}_2(\tau - R_{\eta_2}(t_l^{(2)}))$ for $\tau \in (\tau_i, \tau_{i+1})$. Hence, by (5), it follows that within each piece (τ_i, τ_{i+1}) for $0 \le i \le p$, $\widetilde{\eta}_1 - \widetilde{\eta}_2$ has at most $\mathscr{S}(\mathcal{F})$ sign change points. Also, the point τ_{i+1} itself could be a sign change point in the interval between the last sign change point in piece (τ_i, τ_{i+1}) and the first sign change point in piece (τ_{i+1}, τ_{i+2}) where $0 \le i \le p-1$. These are all the possible sign change points of $\widetilde{\eta}_1 - \widetilde{\eta}_2$. Hence,

$$\begin{aligned} \sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2) &\le p + (p+1)\mathscr{S}(\mathcal{F}) \\ &\le (p_1 + p_2) + (p_1 + p_2 + 1)\mathscr{S}(\mathcal{F}) \\ &\le (2K-2) + (2K-1)\mathscr{S}(\mathcal{F}). \end{aligned} \tag{11}$$

Since (11) holds for all $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$, the lemma follows. □

PROOF OF THEOREM 4. The proof is by induction on the number of sign changes of f. If f has zero sign changes, then without loss of generality, $f(t) \ge 0$

for $t \in (0, \infty)$ and $f(t) > 0$ for some interval $(a, b) \subseteq (0, \infty)$. Hence, $G(x) > 0$ for all x, and the base case holds. Suppose f has $m + 1$ sign change points $t_0, \ldots, t_m$, where $m \geq 0$. Note that $G(x)$ and $F(x) = e^{t_0 x} G(x)$ have the same real-valued roots (with multiplicity) since $e^{t_0 x} > 0$ for all $x \in \mathbb{R}$. $F'(x)$ is given by

$$F'(x) = \frac{d}{dx}\left(\int_0^\infty f(t) e^{-(t-t_0)x}\, dt\right) = \int_0^\infty (t_0 - t) f(t) e^{-(t-t_0)x}\, dt,$$

where the interchange of the differential and integral operators in the second equality is justified by the Leibniz integral rule because f is piecewise continuous over $\mathbb{R}_{\geq 0}$, and both $f(t)e^{-(t-t_0)x}$ and $\frac{d}{dx}(f(t)e^{-(t-t_0)x})$ are jointly continuous over $(p_i, p_{i+1}) \times (-\infty, \infty)$ for each piece (p_i, p_{i+1}) over which f is continuous. Note that the set of sign change points of $(t_0 - t) f(t)$ is $\{t_1, \ldots, t_m\}$. Hence, $(t_0 - t) f(t)$ has only m sign changes. By the induction hypothesis, F' has at most m real-valued roots. By Rolle's theorem, the number of real-valued roots of F is at most one more than the number of real-valued roots of F'. Hence, F has at most $m + 1$ real-valued roots, implying that G has at most $m + 1$ real-valued roots. □

PROOF OF THEOREM 5. Suppose there exist two distinct population size functions $\eta_1, \eta_2 \in \mathcal{F}$ that produce exactly the same c_m for all $2 \leq m \leq n$. From (4), we have that

$$\int_0^\infty \big(\widetilde{\eta}_1(\tau) - \widetilde{\eta}_2(\tau)\big) e^{-\binom{m}{2}\tau}\, d\tau = 0 \tag{12}$$

for $2 \leq m \leq n$. If we define the function $G(x)$ as

$$G(x) = \int_0^\infty \big(\widetilde{\eta}_1(\tau) - \widetilde{\eta}_2(\tau)\big) e^{-x\tau}\, d\tau,$$

then from (12), we see that $\binom{m}{2}$ is a root of $G(x)$ for $2 \leq m \leq n$, and hence, G has at least $n - 1$ roots. Applying Theorem 4 to the piecewise continuous function $\widetilde{\eta}_1 - \widetilde{\eta}_2$, we see that G can have at most $\sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2)$ roots. Taking the supremum over all population size functions η_1 and η_2 in $\mathcal{F}$, we see that G can have at most $\mathscr{S}(\mathcal{F})$ roots. Hence, if $n - 1 > \mathscr{S}(\mathcal{F})$, we get a contradiction. This implies that if $n \geq \mathscr{S}(\mathcal{F}) + 2$, no two distinct population size functions in $\mathcal{F}$ can produce the same $(c_2, \ldots, c_n)$. □

PROOF OF COROLLARY 7. As remarked after Lemma 2, for the constant population size function family $\mathcal{F}_c$, $\mathscr{S}(\mathcal{F}_c) = 0$. Hence, by Theorem 6, if $n \geq 2K$, the map $\mathbf{c} : \mathcal{M}_K(\mathcal{F}_c) \to \mathbb{R}_+^{n-1}$ is injective. □

AN ALTERNATE PROOF OF COROLLARY 7 BASED ON LINEAR ALGEBRA. Let $n \geq 2K$, and suppose there exist two distinct models $\eta^{(1)}, \eta^{(2)} \in \mathcal{M}_K(\mathcal{F}_c)$ that produce exactly the same c_m for all $2 \leq m \leq n$. Let $\widetilde{\eta}^{(1)}$ and $\widetilde{\eta}^{(2)}$ denote the time-rescaled versions of $\eta^{(1)}$ and $\eta^{(2)}$, respectively, as in (2). Since $\eta^{(j)}$ is

piecewise constant with at most K pieces, $\widetilde{\eta}^{(j)}$ is also piecewise constant with the same number of pieces as $\eta^{(j)}$, and $\eta^{(1)} \neq \eta^{(2)}$ implies $\widetilde{\eta}^{(1)} \neq \widetilde{\eta}^{(2)}$. Therefore, $\widetilde{\Delta} := \widetilde{\eta}^{(1)} - \widetilde{\eta}^{(2)}$ is a piecewise-constant function over $[0, \infty)$ with p pieces, where $1 \leq p \leq 2K - 1$, and $\widetilde{\Delta}$ is not identically zero. Let $\tau_1 < \cdots < \tau_{p-1}$ denote the change points of $\widetilde{\Delta}$, and define $\tau_0 = 0$ and $\tau_p = \infty$. Suppose $\widetilde{\Delta}(\tau) = \delta_i \in \mathbb{R}$ for all $\tau \in [\tau_{i-1}, \tau_i)$, where $1 \leq i \leq p$. Since $\widetilde{\eta}^{(1)}$ and $\widetilde{\eta}^{(2)}$ produce the same c_m for all $2 \leq m \leq n$, we know that $\widetilde{\Delta}$ satisfies

$$\int_0^\infty \widetilde{\Delta}(\tau) e^{-\binom{m}{2}\tau}\, d\tau = 0, \tag{13}$$

for all $2 \leq m \leq n$. Substituting the definition of $\widetilde{\Delta}$ into (13) and multiplying by $\binom{m}{2}$, we obtain

$$\sum_{i=1}^{p} \delta_i \big[e^{-\binom{m}{2}\tau_{i-1}} - e^{-\binom{m}{2}\tau_i} \big] = 0, \tag{14}$$

for $2 \leq m \leq n$. This defines a linear system $\mathbf{A}\delta = \mathbf{0}$, where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ and $\mathbf{A} = (a_{mi})$ is an $(n-1) \times p$ matrix with $a_{mi} := e^{-\binom{m}{2}\tau_{i-1}} - e^{-\binom{m}{2}\tau_i}$ for $2 \leq m \leq n$ and $1 \leq i \leq p$.

Let $\mathbf{B} = (b_{mi})$ be the $(n-1) \times p$ matrix formed from $\mathbf{A}$ such that the ith column of $\mathbf{B}$ is the sum of columns $i, i+1, \ldots, p$ of $\mathbf{A}$. Defining $\alpha_i = e^{-\tau_{i-1}}$, note that $b_{mi} = \alpha_i^{\binom{m}{2}}$ for $2 \leq m \leq n$ and $1 \leq i \leq p$. Now, consider the $p \times p$ submatrix $\mathbf{C}$ of $\mathbf{B}$ consisting of the first p rows of $\mathbf{B}$. Since $\alpha_1 > \alpha_2 > \cdots > \alpha_p > 0$, note that $\mathbf{C}$ is a generalized Vandermonde matrix, which implies $\det(\mathbf{C}) \neq 0$ [8], Chapter XIII, Section 8. Hence, $\operatorname{rank}(\mathbf{B}) = p$. The rank of $\mathbf{A}$ is invariant under elementary column operations and, therefore, $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{B}) = p$. Therefore, the kernel of $\mathbf{A}$ is trivial, and the only solution to (14) is $\delta_1 = \delta_2 = \cdots = \delta_p = 0$, which contradicts our assumption that $\widetilde{\Delta} = \widetilde{\eta}^{(1)} - \widetilde{\eta}^{(2)} \not\equiv 0$. □

PROOF OF COROLLARY 8. Let $f_1, f_2 \in \mathcal{F}_e$ be given by

$$f_1(t) = \nu_1 \exp(\beta_1 t),$$
$$f_2(t) = \nu_2 \exp(\beta_2 t),$$

where $t \in \mathbb{R}_{\geq 0}$, $\nu_1, \nu_2 \in \mathbb{R}_+$ and $\beta_1, \beta_2 \in \mathbb{R}$. Then, for $i = 1, 2$, the time-rescaled function $\widetilde{f}_i$ is given by

$$\widetilde{f}_i(\tau) = \frac{\nu_i}{1 - \nu_i \beta_i \tau}, \tag{15}$$

for $\tau \in \operatorname{Dom}(\widetilde{f}_i) = R_{f_i}(\mathbb{R}_{\geq 0}) = [0, \frac{1}{\nu_i \beta_i})$. From (15), it can be seen that $\widetilde{f}_1$ and $\widetilde{f}_2$ are continuous in their domains. Furthermore, for any given $a \in \mathbb{R}_{\geq 0}$, there is at most one τ, where $\tau \in \operatorname{Dom}(\widetilde{f}_1)$ and $\tau - a \in \operatorname{Dom}(\widetilde{f}_2)$, such that $g(\tau) := \widetilde{f}_1(\tau) - \widetilde{f}_2(\tau - a) = 0$, implying $\sigma(g) \leq 1$. By the definition of sign change complexity

in (5), it then follows that $\mathscr{S}(\mathcal{F}_e) \leq 1$ for the exponential population family $\mathcal{F}_e$. Hence, applying Theorem 6, we conclude that $n \geq 4K - 1$ suffices for the map $\mathbf{c}: \mathcal{M}_K(\mathcal{F}_e) \to \mathbb{R}_+^{n-1}$ to be injective. $\square$

PROOF OF COROLLARY 9. Let $f_1, f_2 \in \mathcal{F}_g$ be generalized-exponential functions which satisfy the following differential equations and initial conditions:

$$\frac{df_i}{dt} = \beta_i f_i(t)^{\gamma_i},$$
$$f_i(0) = \nu_i, \qquad i \in \{1, 2\},$$

where $\nu_i \in \mathbb{R}_+$, $\beta_i \in \mathbb{R}$ and $\gamma_i \in \mathbb{R}_{\geq 0}$. The solutions for f_i are given by

$$f_i(t) = \begin{cases} \nu_i \exp(\beta_i t), & \gamma_i = 1, \\ \left[\nu_i^{1-\gamma_i} + \beta_i t(1-\gamma_i)\right]^{1/(1-\gamma_i)}, & \gamma_i \neq 1. \end{cases}$$

It can be shown that the time-rescaled population size functions $\widetilde{f}_i$ are given by

$$\widetilde{f}_i(\tau) = \begin{cases} \nu_i \exp(\beta_i \tau), & \beta_i = 0 \text{ or } \gamma_i = 0, \\ \left(\nu_i^{-\gamma_i} - \beta_i \gamma_i \tau\right)^{-1/\gamma_i}, & \beta_i \neq 0 \text{ and } \gamma_i > 0. \end{cases} \tag{16}$$

In order to obtain an upper bound on $\mathscr{S}(\mathcal{F}_g)$, we consider the following three cases depending on the functional form of $\widetilde{f}_1$ and $\widetilde{f}_2$ in (16):

- *Case* 1: $\widetilde{f}_1(\tau) = \nu_1 \exp(\beta_1 \tau)$ and $\widetilde{f}_2(\tau) = \nu_2 \exp(\beta_2 \tau)$. Since $\widetilde{f}_1$ and $\widetilde{f}_2$ are continuous functions of τ, the number of sign changes of $g(\tau) := \widetilde{f}_1(\tau) - \widetilde{f}_2(\tau - a)$ is at most the number of roots of $g(\tau)$. Taking the logarithm of $\widetilde{f}_1(\tau)$ and $\widetilde{f}_2(\tau - a)$, it is easy to see that $g(\tau)$ has at most one root for any $a \in \mathbb{R}_{\geq 0}$. Hence, $\sigma(g) \leq 1$.
- *Case* 2: $\widetilde{f}_1$ and $\widetilde{f}_2$ have different functional forms. Suppose $\widetilde{f}_1(\tau) = \nu_1 \exp(\beta_1 \tau)$ and $\widetilde{f}_2(\tau) = (\nu_2^{-\gamma_2} - \beta_2 \gamma_2 \tau)^{-1/\gamma_2}$. For any $a_1, a_2 \in \mathbb{R}_{\geq 0}$ such that $\tau - a_i \in \mathrm{Dom}(\widetilde{f}_i)$, the number of sign changes of $g(\tau) := \widetilde{f}_1(\tau - a_1) - \widetilde{f}_2(\tau - a_2)$ is at most the number of roots of $g(\tau)$. By raising $\widetilde{f}_1(\tau - a_1)$ and $\widetilde{f}_2(\tau - a_2)$ to the power of $-\gamma_2$, we see that the number of roots of $g(\tau)$ is the number of solutions to

$$\mu_1 \exp(-\gamma_2 \beta_1 \tau) = \mu_2^{-\gamma_2} - \beta_2 \gamma_2 \tau, \tag{17}$$

  where $\mu_1 = \nu_1 \exp(\gamma_2 \beta_1 a_1)$ and $\mu_2 = (\nu_2^{-\gamma_2} + \beta_2 \gamma_2 a_2)^{-1/\gamma_2}$. Equation (17) represents the intersection of an exponential function with a line and has at most 2 solutions for τ. Hence, $\sigma(g) \leq 2$.
- *Case* 3: $\widetilde{f}_i(\tau) = (\nu_i^{-\gamma_i} - \beta_i \gamma_i \tau)^{-1/\gamma_i}$ for $i = 1, 2$. Let $g(\tau) := \widetilde{f}_1(\tau) - \widetilde{f}_2(\tau - a)$ where $a \in \mathbb{R}_{\geq 0}$ such that $\tau - a \in \mathrm{Dom}(\widetilde{f}_2)$. Since g is a continuous function, the number of sign changes of $g(\tau)$ in $\mathbb{R}_{\geq 0}$ is bounded by the number of distinct positive roots of $g(\tau)$. The number of distinct positive roots of g is the number of distinct positive solutions τ to

$$\left(\nu_1^{-\gamma_1} - \beta_1 \gamma_1 \tau\right)^{-1/\gamma_1} = \left(\nu_2^{-\gamma_2} - \beta_2 \gamma_2 (\tau - a)\right)^{-1/\gamma_2},$$

which is also the number of distinct positive solutions to

$$(\nu_1^{-\gamma_1} - \beta_1\gamma_1\tau)^{\gamma_2/\gamma_1} = \nu_2^{-\gamma_2} - \beta_2\gamma_2(\tau - a). \tag{18}$$

Let $x := \widetilde{f}_1(\tau)^{-\gamma_1} = \nu_1^{-\gamma_1} - \beta_1\gamma_1\tau$. Since $\widetilde{f}_1$ is a time-rescaled population size function, $x > 0$ when $\tau \in \mathbb{R}_{\geq 0}$. Since $\beta_i \neq 0$ and $\gamma_i > 0$, (18) can be rewritten as

$$x^{\gamma_2/\gamma_1} + Ax + B = 0,$$

where $A = -\frac{\beta_2\gamma_2}{\beta_1\gamma_1}$ and $B = \frac{\beta_2\gamma_2}{\beta_1\gamma_1}\nu_1^{-\gamma_1} - \nu_2^{-\gamma_2} - \beta_2\gamma_2 a$. Letting $h(x) := x^{\gamma_2/\gamma_1} + Ax + B$, the number of distinct positive solutions for τ in (18) is at most the number of distinct positive roots for the generalized polynomial h. For any real-valued function $g(x)$ possessing infinitely many derivatives and any interval $I \subseteq \mathbb{R}$, let $Z(g, I)$ the number of zeroes of g contained in I, counted with multiplicity. By a consequence of Rolle's theorem [15], Proposition 2.1, $Z(g, I) \leq Z(g', I) + 1$. Observing that $h'(x) = \frac{\gamma_2}{\gamma_1}x^{\gamma_2/\gamma_1 - 1} + A$ has at most one root in $\mathbb{R}_+$, $Z(h, \mathbb{R}_+) \leq Z(h', \mathbb{R}_+) + 1 \leq 2$. Hence, the number of distinct positive solutions τ to (18) is at most 2, and $\sigma(g) \leq 2$.

From the definition of sign change complexity in (5) and the bound on $\sigma(g)$ in the three cases above, it follows that $\mathscr{S}(\mathcal{F}_g) \leq 2$ for the generalized-exponential population family $\mathcal{F}_g$. Hence, applying Theorem 6, we conclude that $n \geq 6K - 2$ suffices for the map $\mathbf{c}: \mathcal{M}_K(\mathcal{F}_g) \to \mathbb{R}_+^{n-1}$ to be injective. □

PROOF OF LEMMA 10. For the population size function $\zeta(t)$ defined by $\zeta(t) = \kappa\eta(t/\kappa)$, note that $R_\zeta(t)$ is given by

$$R_\zeta(t) = \int_0^t \frac{1}{\zeta(x)}\,dx = \int_0^t \frac{1}{\kappa\eta(x/\kappa)}\,dx = \int_0^{t/\kappa} \frac{1}{\eta(x)}\,dx = R_\eta(t/\kappa).$$

$\mathbb{E}[T_{m,m}^{(\zeta)}]$ is then given by

$$\begin{aligned}\mathbb{E}[T_{m,m}^{(\zeta)}] &= \int_0^\infty \exp\left[-\binom{m}{2} R_\eta\left(\frac{t}{\kappa}\right)\right] dt \\ &= \kappa \int_0^\infty \exp\left[-\binom{m}{2} R_\eta(t)\right] dt \\ &= \kappa\mathbb{E}[T_{m,m}^{(\eta)}].\end{aligned}$$

Since $n \geq 2K + (2K - 1)\mathscr{S}(\mathcal{F})$, by Theorem 6, ζ is the unique population size function in $\mathcal{M}_K(\mathcal{F})$ with $\mathbb{E}[T_{m,m}^{(\zeta)}] = \kappa c_m$ for $2 \leq m \leq n$. □

To prove Theorem 12, we first need a lemma that characterizes a certain symmetry property of the invertible matrix that relates the genealogical quantities $\boldsymbol{\gamma}$ and $\mathbf{c}$ introduced in the proof of Lemma 1.

LEMMA 13. *For a sample of size n, let W be the $(n-1)\times(n-1)$ invertible matrix such that $\gamma_{n,b} = \sum_{m=2}^{n} W_{b,m}c_m$, where $\gamma_{n,b}$ is the total expected branch length subtending b leaves and $c_m = \mathbb{E}[T^{(\eta)}_{m,m}]$. Then, for every b and m, where $1 \le b \le n-1$ and $2 \le m \le n$, we have the following identities*:

$$W_{b,m} + W_{n-b,m} = 0 \qquad \text{if } m \text{ is odd},$$
$$W_{b,m} - W_{n-b,m} = 0 \qquad \text{if } m \text{ is even}.$$

PROOF. From the proof of Lemma 1, it can be seen that the matrix $\mathbf{W}$ is the product of 3 matrices whose entries are explicitly given combinatorial expressions. However, using Zeilberger's algorithm [34], Polanski and Kimmel [36], equations (13)–(15), also derived the following recurrence relation for the entries of $\mathbf{W}$:

$$\begin{aligned} W_{b,2} &= \frac{6}{(n+1)}, \\ W_{b,3} &= 30\frac{(n-2b)}{(n+1)(n+2)}, \\ W_{b,m+2} &= f(n,m)W_{b,m} + g(n,m)(n-2b)W_{b,m+1}, \end{aligned} \tag{19}$$

where $f(n,m)$ and $g(n,m)$ are rational functions of n and m given by

$$f(n,m) = -\frac{(1+m)(3+2m)(n-m)}{m(2m-1)(n+m+1)},$$
$$g(n,m) = \frac{(3+2m)}{m(n+m+1)}.$$

It will be easy to prove our lemma by induction on m using (19). The base cases are easy to check:

$$W_{b,2} - W_{n-b,2} = 0,$$
$$W_{b,3} + W_{n-b,3} = 30\frac{(n-2b)+(n-2(n-b))}{(n+1)(n+2)} = 0.$$

Using (19), we see that if m is odd,

$$\begin{aligned} &W_{b,m+2} + W_{n-b,m+2} \\ &\quad = f(n,m)(W_{b,m} + W_{n-b,m}) \\ &\qquad + g(n,m)\{(n-2b)W_{b,m+1} + [n-2(n-b)]W_{n-b,m+1}\} \\ &\quad = f(n,m)(W_{b,m} + W_{n-b,m}) + g(n,m)(n-2b)(W_{b,m+1} - W_{n-b,m+1}) \\ &\quad = 0, \end{aligned}$$

where the last equality follows from the induction hypothesis which implies $W_{b,m} + W_{n-b,m} = 0$ and $W_{b,m+1} - W_{n-b,m+1} = 0$. Similarly, if m is even,

$$\begin{aligned} W_{b,m+2} &- W_{n-b,m+2} \\ &= f(n,m)(W_{b,m} - W_{n-b,m}) \\ &\quad + g(n,m)\{(n-2b)W_{b,m+1} - [n-2(n-b)]W_{n-b,m+1}\} \\ &= f(n,m)(W_{b,m} - W_{n-b,m}) + g(n,m)(n-2b)(W_{b,m+1} + W_{n-b,m+1}) \\ &= 0, \end{aligned}$$

where again the last equality follows from the induction hypothesis. □

PROOF OF THEOREM 12. For a sample of size n in the coalescent, let $\gamma_{n,b}$ be the total expected branch length subtending b leaves, for $1 \le b \le n-1$. Then there exists a positive constant κ such that

$$\frac{\gamma_{n,d} + \gamma_{n,n-d}}{1 + \delta_{d,n-d}} = \kappa \chi_{n,d}, \tag{20}$$

for all $1 \le d \le \lfloor n/2 \rfloor$. Let $f_{n,d} = \frac{\gamma_{n,d}+\gamma_{n,n-d}}{1+\delta_{d,n-d}}$. The relationship between $\mathbf{f} = (f_{n,1}, \ldots, f_{n,\lfloor n/2 \rfloor})$ and $\boldsymbol{\gamma} = (\gamma_{n,1}, \ldots, \gamma_{n,n-1})$ can be described by the linear equation

$$\mathbf{f} = \mathbf{Z}\boldsymbol{\gamma},$$

where $\mathbf{Z}$ is an $\lfloor n/2 \rfloor \times (n-1)$ matrix with entries given by

$$Z_{dj} = \begin{cases} 1, & \text{if } j = d \text{ or } j = n-d, \\ 0, & \text{otherwise,} \end{cases}$$

for $1 \le d \le \lfloor n/2 \rfloor$ and $1 \le j \le n-1$. Hence, $\dim(\ker(\mathbf{Z})) = \lfloor (n-1)/2 \rfloor$.

From Lemma 1, we know that $\boldsymbol{\gamma}$ and $\mathbf{c} = (c_2, \ldots, c_n)$ are related as $\boldsymbol{\gamma} = \mathbf{W}\mathbf{c}$, where $\mathbf{W} = (W_{b,m})$ is an $(n-1) \times (n-1)$ invertible matrix, where $1 \le b \le n-1$ and $2 \le m \le n$. Hence,

$$\mathbf{f} = \mathbf{Y}\mathbf{c}, \tag{21}$$

where $\mathbf{Y} := \mathbf{Z}\mathbf{W}$. Since $Y_{b,m} = W_{b,m} + W_{n-b,m}$, we know from Lemma 13 that $Y_{b,m} = 0$ for all odd values of m. Therefore, every other column of the matrix $\mathbf{Y}$ is zero. This implies that $\operatorname{span}(\{\mathbf{e}_3, \mathbf{e}_5, \ldots, \mathbf{e}_{n-\mathbb{1}\{n \text{ even}\}}\}) \subseteq \ker(\mathbf{Y})$, where $\mathbf{e}_i$ is an $(n-1)$-dimensional unit vector defined as $\mathbf{e}_i = (e_{i,2}, \ldots, e_{i,n})$, with $e_{i,i} = 1$ and $e_{i,j} = 0$ for $i \ne j$. Note that $n - \mathbb{1}\{n \text{ even}\} = 2\lfloor (n-1)/2 \rfloor + 1$ and $\dim(\operatorname{span}(\{\mathbf{e}_3, \mathbf{e}_5, \ldots, \mathbf{e}_{2\lfloor (n-1)/2 \rfloor + 1}\})) = \lfloor (n-1)/2 \rfloor$. Now, since $\mathbf{W}$ is invertible, $\dim(\ker(\mathbf{Y})) = \dim(\ker(\mathbf{Z}\mathbf{W})) = \dim(\ker(\mathbf{Z})) = \lfloor (n-1)/2 \rfloor$. Therefore,

$$\ker(\mathbf{Y}) = \operatorname{span}\big(\{\mathbf{e}_3, \mathbf{e}_5, \ldots, \mathbf{e}_{2\lfloor (n-1)/2 \rfloor + 1}\}\big). \tag{22}$$

Suppose there exist two distinct models $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$ that produce the same folded SFS $\mathbf{f}$. Let $\mathbf{c}^{(1)}$ and $\mathbf{c}^{(2)}$ be the vector of genealogical quantities for models η_1 and η_2, respectively, where $c_m^{(1)} = \mathbb{E}[T_{m,m}^{(\eta_1)}]$ and $c_m^{(2)} = \mathbb{E}[T_{m,m}^{(\eta_2)}]$, $2 \le m \le n$. From (21), we know that $\mathbf{c}^{(1)} - \mathbf{c}^{(2)} \in \ker(\mathbf{Y})$. Using (22), $c_m^{(1)} - c_m^{(2)}$ can be written as

$$c_m^{(1)} - c_m^{(2)} = \sum_{l=1}^{\lfloor (n-1)/2 \rfloor} \alpha_l e_{2l+1,m}, \tag{23}$$

for some $\alpha_l \in \mathbb{R}$. Since $e_{ij} = 0$ for $i \neq j$, (23) implies that $c_m^{(1)} - c_m^{(2)} = 0$ for all even values of m, where $2 \le m \le n$. Now applying a similar argument as in the proof of Theorem 6 to $c_m^{(1)} - c_m^{(2)}$ for even values of m, we conclude that if $\lceil (n-1)/2 \rceil > (2K-2) + (2K-1)\mathscr{S}(\mathcal{F})$, then no two distinct models $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$ can produce the same $\mathbf{f}$. This implies that a sample size $n \ge 2(2K-1)(1+\mathscr{S}(\mathcal{F}))$ suffices for identifying the population size function in $\mathcal{M}_K(\mathcal{F})$ from the folded SFS $\mathbf{f}$, and the conclusion of the theorem follows from (20) and Lemma 10. □

4. Discussion. In human genetics, several large-sample datasets have recently become available, with sample sizes on the order of several thousands to tens of thousands of individuals [1, 4, 6, 31, 44]. The patterns of polymorphism observed in these datasets deviate significantly from that expected under a constant population size, and there has been much interest in inferring recent and ancient human demographic changes that might explain these deviations [10, 24, 26]. Clearly, model identifiability is an important prerequisite for such statistical inference problems. In this paper, we have obtained mathematically rigorous identifiability results for demographic inference by showing that piecewise-defined population size functions over a wide class of function families are completely determined by the SFS, provided that the sample is sufficiently large. Furthermore, we have provided explicit bounds on the sample sizes that are sufficient for identifying such piecewise population size functions. These bounds depend on the number of pieces and the functional type of each piece. For piecewise-constant population size models, which have been extensively applied in demographic inference studies, our bounds are tight. We have also given analogous results for identifiability from the folded SFS, a variant of the SFS that is oblivious to the identities of the ancestral and mutant alleles.

Recent large-sample sequencing studies have consistently found a substantially higher fraction of rare variants compared to the predictions of the coalescent with a constant population size, even in regions of the genome that are believed to have evolved neutrally [9]. Keinan and Clark [16] suggested that recent rapid expansion of the population has given rise to variants which are private to single individuals in the population, and that this signature of population expansion is particularly

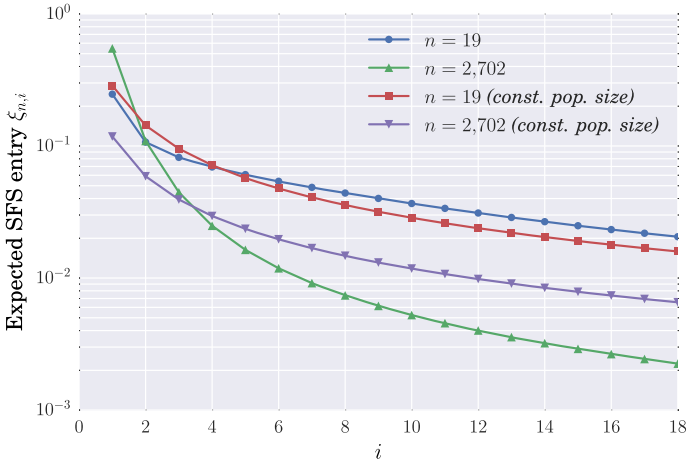

FIG. 3. *The leading entries of the expected SFS* $\boldsymbol{\xi}_n$ *for a piecewise-exponential population size model inferred b Tennessen et al.* [44]. *This demographic model, shown (up to scaling) in Figure* 1, *was fitted using the observed SFS from a sample of* 1351 *(diploid) individuals of European ancestry* [44]. *The blue plot is the expected SFS for* $n = 19$, *which matches the sample size bound in Corollary* 8 *for identifying piecewise-exponential models with up to* 5 *pieces, while the green plot is the first* 18 *entries of the expected SFS for* $n = 2702$ (1351 *diploids*). *The red and purple plots are the expected SFS for* $n = 19$ *and* $n = 2702$, *respectively, for a constant population size function.*

apparent now due to the larger sample sizes involved in sequencing studies. We illustrate this point with a specific example. The blue plot in Figure 3 shows the expected SFS for a sample of size $n = 19$ under the piecewise-exponential population size history with 5 epochs recently inferred by Tennessen et al. [44] and illustrated in Figure 1. (Note that $n = 19$ is the sample size bound given by Corollary 8 for identifying piecewise-exponential models with up to 5 pieces.) The red plot in Figure 3 shows the expected SFS for the same sample size under a constant population size model. For this small sample size, the two expected frequency spectra are very similar despite the large difference in demographic models, indicating the difficulty of accurately recovering the details of recent exponential population growth using small-sample data. In contrast, for a much larger sample of size $n = 2702$, which corresponds to the actual sample size for Tennessen et al.'s data, the expected frequency spectra under the two demographic models mentioned above are considerably more different; see the green and purple plots in Figure 3.

On the other hand, our identifiability results show that *perfect* data (i.e., the *exact* expected SFS) from even a small sample size of $n = 4K - 1$ are sufficient to uniquely identify a piecewise-exponential model with K pieces. This gap between theoretical identifiability and practical inference needs to be better addressed through robustness results that can account for the finite genome length, which limits the resolution to which the expected SFS of a random sample can be

estimated. Our identifiability results apply in the limit that the genome length is infinite, which allows one to estimate the entries of the expected SFS exactly. On the other hand, a finite length genome does not permit exact estimation of the expected SFS, which can make it difficult in practice to resolve the details of ancient demographic events even if the sample size is large. This is because population size changes sufficiently far back in the past are likely to have only a marginal effect on the SFS since the individuals in the sample are highly likely to have found a common ancestor by such ancient times.

Our work suggests several interesting avenues for future research. An important problem is to understand the sensitivity of the SFS to perturbations in the demographic parameters. A related problem is quantifying the extent to which errors in estimating the expected SFS from a finite amount of data affect the parameter estimates in inferred demographic models.

It would also be interesting to consider the possibility of developing an algebraic algorithm for demographic inference that closely mimics the linear algebraic proof of Corollary 7 provided in Section 3. For example, using a sample of size $K + 1$, one could consider inferring a piecewise-constant model with K pieces, with one piece for each of the most recent $K - 1$ generations and another piece for the population size further back in time. (Here, we are considering a restricted class of piecewise-constant population size functions with fixed change points, so the minimum sample size needed for distinguishing such models using the SFS is $K + 1$ rather than $2K$.) Such an algebraic algorithm could provide a more principled way of inferring demographic parameters, compared to existing inference methods that rely on optimization procedures which lack theoretical guarantees for functions with multiple local optima.

In our work, we focused on the identifiability of demography from the expected SFS data. However, if one were to use the complete sequence data or other summary statistics such as the length distribution of shared haplotype tracts, it might be possible to uniquely identify the demography using even smaller sample sizes than that needed when using only the SFS. Indeed, several demographic inference methods have been developed to infer historical population size changes from such data using anywhere from a pair of genomic sequences [14, 24, 32] to tens of such sequences [42], and it is important to theoretically characterize the power and limitations of both the data and the inference methods.

Acknowledgments. We thank Graham Coop, Noah Mattoon, Aylwyn Scally and anonymous reviewers for their comments on our work. We also thank the generous support of the Simons Institute for the Theory of Computing. The final version of this work was completed while the authors were participating in the 2014 program on "Evolutionary Biology and the Theory of Computing."

REFERENCES

[1] 1000 GENOMES PROJECT CONSORTIUM (2010). A map of human genome variation from population-scale sequencing. *Nature* **467** 1061–1073.

[2] BOYKO, A. R., WILLIAMSON, S. H., INDAP, A. R., DEGENHARDT, J. D., HERNANDEZ, R. D., LOHMUELLER, K. E., ADAMS, M. D., SCHMIDT, S., SNINSKY, J. J., SUNYAEV, S. R. et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*. **4** e1000083.

[3] CAMPBELL, C. D., OGBURN, E. L., LUNETTA, K. L., LYON, H. N., FREEDMAN, M. L., GROOP, L. C., ALTSHULER, D., ARDLIE, K. G. and HIRSCHHORN, J. N. (2005). Demonstrating stratification in a European American population. *Nat. Genet*. **37** 868–872.

[4] COVENTRY, A., BULL-OTTERSON, L. M., LIU, X., CLARK, A. G., MAXWELL, T. J., CROSBY, J., HIXSON, J. E., REA, T. J., MUZNY, D. M., LEWIS, L. R. et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications* **1** 131.

[5] EXCOFFIER, L., DUPANLOUP, I., HUERTA-SÁNCHEZ, E., SOUSA, V. C. and FOLL, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet*. **9** e1003905.

[6] FU, W., O'CONNOR, T. D., JUN, G., KANG, H. M., ABECASIS, G., LEAL, S. M., GABRIEL, S., ALTSHULER, D., SHENDURE, J., NICKERSON, D. A. et al. (2012). Analysis of 6515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493** 216–220.

[7] FU, Y. X. (1995). Statistical properties of segregating sites. *Theor. Popul. Biol*. **48** 172–197.

[8] GANTMACHER, F. R. (2000). *The Theory of Matrices*, *Vol*. 2. Chelsea, New York.

[9] GAZAVE, E., CHANG, D., CLARK, A. G. and KEINAN, A. (2013). Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics* **195** 969–978.

[10] GRAVEL, S., HENN, B. M., GUTENKUNST, R. N., INDAP, A. R., MARTH, G. T., CLARK, A. G., YU, F., GIBBS, R. A., BUSTAMANTE, C. D., ALTSHULER, D. L. et al. (2011). Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. USA* **108** 11983–11988.

[11] GRIFFITHS, R. C. (2003). The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theor. Popul. Biol*. **64** 241–251.

[12] GRIFFITHS, R. C. and TAVARÉ, S. (1998). The age of a mutation in a general coalescent tree. *Comm. Statist. Stochastic Models* **14** 273–295. MR1617552

[13] GUTENKUNST, R. N., HERNANDEZ, R. D., WILLIAMSON, S. H. and BUSTAMANTE, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. **5** e1000695.

[14] HARRIS, K. and NIELSEN, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*. **9** e1003521.

[15] JAMESON, G. J. O. (2006). Counting zeros of generalised polynomials: Descartes' rule of signs and Laguerre's extensions. *The Mathematical Gazette* **90** 223–234.

[16] KEINAN, A. and CLARK, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336** 740–743.

[17] KEINAN, A., MULLIKIN, J. C., PATTERSON, N. and REICH, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East asians than in europeans. *Nat. Genet*. **39** 1251–1255.

[18] KIDD, J. M., GRAVEL, S., BYRNES, J., MORENO-ESTRADA, A., MUSHAROFF, S., BRYC, K., DEGENHARDT, J. D., BRISBIN, A., SHETH, V., CHEN, R. et al. (2012). Population genetic inference from personal genome data: Impact of ancestry and admixture on human genomic variation. *Am. J. Hum. Genet*. **91** 660–671.

[19] KIMURA, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA* **41** 144–150.

[20] KIMURA, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61** 893.
[21] KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR0671034
[22] KINGMAN, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probab.* **19A** 27–43. MR0633178
[23] KINGMAN, J. F. C. (1982). Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics* (*Rome*, 1981) (G. Koch and F. Spizzichino, eds.) 97–112. North-Holland, Amsterdam. MR0675968
[24] LI, H. and DURBIN, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475** 493–496.
[25] LOHMUELLER, K. E., INDAP, A. R., SCHMIDT, S., BOYKO, A. R., HERNANDEZ, R. D., HUBISZ, M. J., SNINSKY, J. J., WHITE, T. J., SUNYAEV, S. R., NIELSEN, R. et al. (2008). Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451** 994–997.
[26] LUKIC, S. and HEY, J. (2012). Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* **192** 619–639.
[27] LUKIĆ, S., HEY, J. and CHEN, K. (2011). Non-equilibrium allele frequency spectra via spectral methods. *Theor. Popul. Biol.* **79** 203–219.
[28] MARCHINI, J., CARDON, L. R., PHILLIPS, M. S. and DONNELLY, P. (2004). The effects of human population structure on large genetic association studies. *Nat. Genet.* **36** 512–517.
[29] MARTH, G. T., CZABARKA, E., MURVAI, J. and SHERRY, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166** 351–372.
[30] MYERS, S., FEFFERMAN, C. and PATTERSON, N. (2008). Can one learn history from the allelic spectrum? *Theor. Popul. Biol.* **73** 342–348.
[31] NELSON, M. R., WEGMANN, D., EHM, M. G., KESSNER, D., JEAN, P. S., VERZILLI, C., SHEN, J., TANG, Z., BACANU, S.-A., FRASER, D. et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337** 100–104.
[32] PALAMARA, P. F., LENCZ, T., DARVASI, A. and PE'ER, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91** 809–822.
[33] PASANIUC, B., ZAITLEN, N., LETTRE, G., CHEN, G. K., TANDON, A., KAO, W. L., RUCZINSKI, I., FORNAGE, M., SISCOVICK, D. S., ZHU, X. et al. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet.* **7** e1001371.
[34] PETKOVŠEK, M., WILF, H. S. and ZEILBERGER, D. (1996). $A = B$. A K Peters, Wellesley, MA. MR1379802
[35] POLANSKI, A., BOBROWSKI, A. and KIMMEL, M. (2003). A note on distributions of times to coalescence, under time-dependent population size. *Theor. Popul. Biol.* **63** 33–40.
[36] POLANSKI, A. and KIMMEL, M. (2003). New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165** 427–436.
[37] PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.
[38] REPPELL, M., BOEHNKE, M. and ZÖLLNER, S. (2012). FTEC: A coalescent simulator for modeling faster than exponential growth. *Bioinformatics* **28** 1282–1283.
[39] REPPELL, M., BOEHNKE, M. and ZÖLLNER, S. (2014). The impact of accelerating faster than exponential population growth on genetic variation. *Genetics* **196** 819–828.

[40] SANKARARAMAN, S., PATTERSON, N., LI, H., PÄÄBO, S. and REICH, D. (2012). The date of interbreeding between Neandertals and modern humans. *PLoS Genet*. **8** e1002947.

[41] SCHAFFNER, S. F., FOO, C., GABRIEL, S., REICH, D., DALY, M. J. and ALTSHULER, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*. **15** 1576–1583.

[42] SHEEHAN, S., HARRIS, K. and SONG, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics* **194** 647–662.

[43] SKOGLUND, P. and JAKOBSSON, M. (2011). Archaic human ancestry in East Asia. *Proc. Natl. Acad. Sci. USA* **108** 18301–18306.

[44] TENNESSEN, J. A., BIGHAM, A. W., O'CONNOR, T. D., FU, W., KENNY, E. E., GRAVEL, S., MCGEE, S., DO, R., LIU, X., JUN, G. et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337** 64–69.

[45] WILLIAMSON, S. H., HERNANDEZ, R., FLEDEL-ALON, A., ZHU, L., NIELSEN, R. and BUSTAMANTE, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA* **102** 7882–7887.

[46] ŽIVKOVIĆ, D. and STEPHAN, W. (2011). Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theor. Popul. Biol*. **79** 184–191.

COMPUTER SCIENCE DIVISION
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: bhaskar@cs.berkeley.edu

DEPARTMENT OF STATISTICS AND
COMPUTER SCIENCE DIVISION
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: yss@stat.berkeley.edu