An Efficient Computational Approach for Constructing the Allele Frequency Spectrum of Populations with Arbitrary Complex History

Hua Chen[a,b,c,*]

[a] *CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[b] *CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China*
[c] *University of Chinese Academy of Sciences, Beijing 100049, China*

Abstract

The allele frequency spectrum (AFS), or site frequency spectrum, is commonly used to summarize the genomic polymorphism pattern of a sample, which is informative for inferring population history and detecting natural selection. Recently, Chen and Chen (2013) developed a method for analytically deriving the AFS for populations with temporally varying size through the coalescence time-scaling function. However, their approach is only applicable for population history scenarios in which the analytical form of the time-scaling function is tractable. In this paper, we propose a computational approach to extend the method to populations with arbitrary complex history by numerically approximating the time-scaling function. We demonstrate the performance of the approach by constructing the AFS for two population history scenarios: the logistic growth model and the Gompertz growth model, for which the AFS are unavailable with existing approaches.

Keywords: Allele frequency spectrum, complex demography, population history, population genetic inference.

1. Introduction

The allele frequency spectrum (AFS, a.k.a. the site frequency spectrum) is a series of fundamental statistics for summarizing genomic polymorphism. It is defined as the sampling distribution of allele frequencies of genetic polymorphism in a finite sample (Chen, 2012). In practice, the AFS can be the number or proportion of SNPs constructed

[*] Corresponding author: chenh@big.ac.cn

by binning them according the counts of derived alleles. For a sample of n sequences with m identified segregating sites (polymorphic sites), the AFS is written as $\{(S_i), 1 < i < n\}$, with $\sum_{i=1}^{n-1} S_i = m$, where S_i denotes the number of segregating sites in the sample that has i copies of derived alleles among the n haplotypes. The AFS has been a main focus in theoretical and methodological studies in the past decades, since it is informative for the inference of ancient demography of populations (Kimura, 1955). The theoretical expectation of AFS under a given population history and parameter setting can be developed using both coalescent theory and diffusion (Fu, 1995; Griffiths and Tavaré, 1998; Sawyer and Hartl, 1992). Methods for ancestral inference based on the AFS are then developed in a Poisson random field framework by assuming each entry of the AFS follows a Poisson distribution with the mean equal to the theoretical expectation of AFS given a population genetic parameter setting (Sawyer and Hartl, 1992; Bustamante *et al.*, 2001; Fu, 1995; Griffiths and Tavaré, 1998; Wooding *et al.*, 2002; Polanski and Kimmel, 2003; Marth *et al.*, 2004; Williamson *et al.*, 2005; Gutenkunst *et al.*, 2009; Lukić *et al.*, 2011; Živković and Stephan, 2011; Chen, 2012; Excoffier *et al.*, 2013; Gao and Keinan, 2015; Bhaskar *et al.*, 2015; Liu and Fu, 2015). These methods gain popularity with the abundance of genomic sequencing data.

Coalescent theory has been applied to developing the AFS in a single population with time-varying population sizes, including the exponential-growth model (Wooding and Rogers, 2002; Polanski and Kimmel, 2003) and the n-epoch model, which models the population size changes using several consecutive periods (epochs) with different constant sizes (Marth *et al.*, 2004). Compared with the AFS developed with diffusion, the coalescent-based AFS has the advantage of being in analytical form, and the estimation is fast and accurate for small samples. In contrast, the diffusion approximation has to rely on numerical methods, such as finite difference approaches, to approximate the solutions (Williamson *et al.*, 2005; Evans *et al.*, 2007). The coalescent-based AFS is thus very useful in the inference of past demographic history and has been extensively applied to data analysis (Marth *et al.*, 2003; Keinan *et al.*, 2007; Gravel *et al.*, 2011; Gazave *et al.*, 2014).

One limitation of the coalescent-based AFS methods is that we can only analytically derive the AFS for some simple population growth models, such as the n-epoch model and the exponential-growth model or their combinations (Polanski and Kimmel, 2003; Marth *et al.*, 2004; Gazave *et al.*, 2014), and generalization to other complex population histories is often impracticable (Chen, 2012, 2013; Polanski and Kimmel, 2003). A second limitation is that for large samples (e.g., haplotype number $n > 50$), it is hard to accurately calculate the expected AFS from the formulae. The expected coalescence times $\mathbb{E}T_i, 1 \leq i < n$ are essential for deriving the coalescent-based AFS, which contain coefficients in the alternating sum of the hypergeometric series and are explosively large, causing overflow for large sample sizes (Polanski *et al.*, 2003). When the sample size is large, the AFS and its derived statistics are informative for inferring recent population history. And thus, calculating the AFS for large samples becomes common in population genetic inference from genomic data (Coventry *et al.*, 2010; Gazave *et al.*, 2014; Chen *et al.*, 2015). A high-precision arithmetic library is usually adopted to obtain accurate numerical values when analyzing larger samples, which requires tedious programming and intensive computation (Marth *et al.*, 2004). Some alternative solutions were proposed, specifically for the AFS of a single population, e.g., Polanski and Kimmel (2003) replaced it with hypergeometric summation to avoid estimating the coefficients with large values. Their approach can efficiently solve the numerical issue, but it is difficult to generalize this approach to other scenarios with complex population histories for which the integral function in the hypergeometric summation is difficult to compute. Most studies have adopted coalescent simulations to generate a large samples to approximate the AFS under specific demographic histories and applied them to analyzing genomic polymorphism. However, this approach is computationally very intensive (Hudson, 2002; Coventry *et al.*, 2010; Nelson *et al.*, 2012; Excoffier *et al.*, 2013; Gazave *et al.*, 2014; Tennessen *et al.*, 2012).

To address the numerical issue in large samples, Chen and Chen (2013) used the large-sample asymptotic distributions of coalescence times. Griffiths (1984) proved that the coalescence times and ancestral lineage numbers asymptotically follow a normal distribution in a constant population. Chen and Chen (2013) extended their forms to populations with time-varying sizes by using a time-scaling function scheme (see the “Coalescence times” section below; Griffiths and Tavaré (1994); Donnelly and Tavaré (1995); Nordborg (2001)) and then used the first-order Taylor expansion approximation to achieve the coalescence times (and, further, the AFS). They illustrated the usage of this approach by deriving a simple-form formula for the AFS in populations under exponential growth, which shows high accuracy compared with simulated results. Note that the first-order Taylor expansion approximation and time-scaling function approach of Chen and Chen (2013) works for both large and small size samples. Technically, their approach allows them to derive AFS in any populations with arbitrary complex demography. However, as we illustrate in the "Methods" section, for some complex demography, it is difficult to derive the analytical form of the time-scaling function and/or its inverse function, which are essential in deriving the coalescent-based AFS. In this paper, we propose a computational approach to efficiently approximate the analytical formula of the time-scaling function with a finite sum approximation, and find the set of coalescence times $\mathbb{E}T_i, 1 \leq i < n$, with the computing time being nearly constant as the sample size increases. It is applicable to any arbitrary complex history for which the time-scaling function is not tractable. This greatly extends the application of AFS-based methods in population genetic inference and other studies, e.g., cancer evolution. We demonstrate the performance of the approach by obtaining the AFS for two population history scenarios that were difficult to derive using the existing approaches: the logistic growth model and the Gompertz model.

In the following sections we first review the coalescent theory framework for obtaining the AFS for a single population. We then summarize the first-order Taylor expansion approximation method for populations with time-varying size proposed by Chen and Chen (2013). We illustrate the idea of the computational approach for constructing the AFS for arbitrary demography, and we further derive the AFS for populations with two demographic histories to demonstrate its performance.

Modeling framework

For a sample of n lineages (haplotypes), the coalescence time T_k is defined as the time when $k+1$ lineages merge into k lineages, and time is measured backward (from the present to the past). The intercoalescence time $W_k = T_{k-1} - T_k$ is the time during which there are k lineages. Following Fu (1995), we say that any of the k branches spanning the intercoalescence time W_k has the branch of size k. We assume an infinitely-many-sites model for mutations, and further mutations occur on branches along the gene genealogy following a Poisson process. The number of mutations occurring at any branch of size k then follows a Poisson distribution with the mean of $\mu k \mathbb{E}(W_k)$, where μ is the point mutation rate. During the bifurcation process in which k lineages increase to n lineages at present, any of these mutations increases the allele account from a single copy to j among the n lineages with the probability (Feller, 2008; Griffiths and Tavaré, 1998):

$$p_{n,k}(j) = \frac{\binom{n-j-1}{k-1}}{\binom{n-1}{k-1}}. \tag{1}$$

Summing over mutations that occur on branches with different sizes, we can obtain the entries for the AFS:

$$\begin{aligned} \mathbb{E}S_j(n) &= \sum_{k=2}^{n} \frac{\binom{n-j-1}{k-2}}{\binom{n-1}{k-1}} \mu \times k\mathbb{E}(W_k) \\ &= \frac{(n-j-1)!(j-1)!}{(n-1)!} \mu \sum_{k=2}^{n} k(k-1) \times \\ &\qquad \binom{n-k}{j-1} \mathbb{E}(W_k), 0 < j < n. \end{aligned} \tag{2}$$

Note that $\mathbb{E}W_j$ is fundamental in the above framework for constructing the AFS. If we can obtain analytical forms for $\mathbb{E}W_j = \mathbb{E}T_{j-1} - \mathbb{E}T_j$ for a population with complex demography, we can obtain the AFS through Equation 2.

Coalescence times

In a constant-size population, the distribution of coalescence times follows that of the standard Kingman's n-coalescent, which are exponential variables with the mean

$$\mu_m = 2(\frac{1}{m} - \frac{1}{n}), 1 \leq m < n, \tag{3}$$

where μ_m is the coalescence time in units of haploid population size N. In addition, the intercoalescence times are mutually independent.

For a population with time-varying size, we denote its population history as $N(t), t \in$

$[0, \infty)$. It is not trivial to derive the distribution or the expectation of coalescence times for a population with time-varying sizes. The joint distribution of coalescence times $(T_m, \ldots, T_{n-1})$ for populations with time-varying size is (Griffiths and Tavaré, 1998)

$$f_{T_m,\ldots,T_{n-1}}(t_m, \ldots, t_{n-1}) = \prod_{k=m}^{n-1} \frac{\binom{k+1}{2}}{N_0\lambda(t_k)} \exp\Big(- \frac{\binom{k+1}{2}}{N_0} \int_{t_{k-1}}^{t_k} \frac{1}{\lambda(u)} du\Big). \tag{4}$$

Polanski *et al.* (2003) derived the marginal probability density function of coalescence times f_{T_m} by expanding an integral transform of the marginal pdf into partial fractions. Another way to derive f_{T_m} is based on the definition of a pure-death process, in the form of a function of the ancestral lineage number, $P(A_n(t) = m)$ (Griffiths, 2006; Chen, 2012). With the marginal distribution of coalescence times derived, Polanski and Kimmel (2003) obtained the AFS for a population under exponential growth, which is in complex form, and requires calculating the hypergeometric series and exponential integral.

Chen and Chen (2013) used the time rescaling approach in the variable-population-size coalescent model (Griffiths and Tavaré, 1994; Nordborg, 2001; Donnelly and Tavaré, 1995). The coalescence time is rescaled at the rate $1/N(t)$, denoted as τ_m:

$$\tau_m = g(T_m) = \int_0^{T_m} \frac{1}{N(u)} du. \tag{5}$$

τ_m follows the coalescence time distribution in the standard Kingman's n-coalescent (Kingman, 1982). Chen and Chen (2013) then used a Taylor expansion of $T_m = g^{-1}(\tau_m)$ around μ_m to achieve the approximation:

$$\begin{aligned} T_m \quad = \quad & g^{-1}(\mu_m) + (g^{-1})'(\mu_m)(g(T_m) - \mu_m) \\ & + \frac{(g^{-1})''(\mu_m)}{2}(g(T_m) - \mu_m)^2 + O((g(T_m) - \mu_m)^3). \end{aligned} \tag{6}$$

Thus

$$\mathbb{E}(T_m) \approx g^{-1}(\mu_m), \tag{7}$$

and

$$Var(T_m) = \frac{\sigma_m^2}{(g'(g^{-1}(\mu_m)))^2}. \tag{8}$$

In general, for any population history $N(t), 0 \leq t < \infty$, we can always obtain the time-scaling function $g(t)$ as in Equation 5, and further obtain $\mathbb{E}T_m = g^{-1}(\mu_m)$ as above. Chen and Chen (2013) demonstrate the application of this approach using an exponentially growing population as an example. $\mathbb{E}T$ for the exponential growth model is in a very simple analytical form:

$$\mathbb{E}T_m = \frac{1}{\gamma}\ln(2N_0\gamma(1/m - 1/n) + 1), \tag{9}$$

and the obtained AFS is highly accurate (see Figure 6 of Chen and Chen (2013)).

Since it is not trivial to derive the coalescence times for populations with time-varying size in existing studies, and simulations are usually required as a replacement for most studies, Chen and Chen (2013)'s approach provides simple and efficient solution for obtaining $\mathbb{E}T_m$ (Coventry *et al.*, 2010; Nelson *et al.*, 2012; Tennessen *et al.*, 2012; Excoffier *et al.*, 2013; Gazave *et al.*, 2014). However, for some complex demographies, the analytical form of the time-scaling function $g(t)$ and its inverse function, which are essential for deriving $\mathbb{E}T_m$, are not tractable. This prohibits the general usage of their approach for arbitrary population history.

Coalescence times under complex demographic history

In this section, we illustrate how to extend Chen and Chen (2013)'s method to be applicable to arbitrary population history using a computational approach. As we can see from the above section, $g(t)$ and $g^{-1}(t)$ are the two essential components for deriving coalescence times for a given population history $N(t)$ (see Equation 7). Note that to obtain $\mathbb{E}T_m$, we do not need the analytical form for calculating an arbitrary point t. In contrast, we only need to find a finite number of T_m values that correspond to $\mu_m, 1 \leq m < n$ and satisfy

$$\mu_m = g(T_m). \tag{10}$$

We thus propose the following two numerical schemes for calculating $\mathbb{E}T_m$, applicable to different situations. The first approach is generally applicable to all cases, including those for which we cannot obtain $g(t)$; the second approach is specifically for the case in which we have an analytical form of $g(t)$ but $g^{-1}(t)$ is not tractable.

Approach 1 (finite sum approximation)

```
Algorithm: Calculating coalescence times

Input: population history $N(t), 0 \leq t < \infty$, sample size $n$.
Initialize: $\mu_i = 2(\frac{1}{i-1} - \frac{1}{n}), i = 1, 2, ..., n-1$; $t = 0$; $G = \frac{1}{N(0)}$.
For $i = n-1 : 1$
    $\mu = \mu_i$;
    While $G < \mu$
        $t = t + 1$;
        $G = G + \frac{1}{N(t)}$;
    End
    If $G - \mu < \frac{1}{2N(t)}$
        $\mathbb{E}T_i = t$;
    Else
        $\mathbb{E}T_i = t - 1$;
    End
End
Output: Expected coalescence time $\mathbb{E}T_i, i = 1, 2, ..., n-1$.
```

Table 1: Procedures for calculating coalescence times using the finite-sum approximation (Approach 1).

For a sample of size n under the population history $N(t), t \in [0, \infty)$, we can simply approximate the integral of the time scaling function equation using the discrete finite summation:

$$\begin{aligned}\mu_m = g(T_m) &= \int_0^{T_m} 1/N(u)du, 1 \leq m < n. \\ &\approx \sum_{u=0}^{T_m} \frac{1}{N(u)}. \end{aligned} \tag{11}$$

Then, for each μ_m, the corresponding expected coalescence times $\mathbb{E}T_m$ can be obtained during the following sequential summation procedures:

Step 1 We have a series of expected coalescence times under the standard n-Kingman's coalescent $\mu_m = 2(\frac{1}{m-1} - \frac{1}{n}), 1 \leq m < n$. Initialize the procedure from generation 0 (the current generation) with $G = \frac{1}{N(0)}$.

Step 2 Keep increasing the discrete generation time t, and calculate $G = G + \frac{1}{N(t)}$ until the value t satisfies $\mu_{n-1} \approx \sum_{u=1}^{t} \frac{1}{N(u)}$. Set $T_{n-1} = t$.

Step 3 Repeat Step 2, and keep increasing t to obtain the rest of the values for $\mathbb{E}T_i, n-2 \leq i \leq 1$.

Step 4 terminate the process when $\mathbb{E}T_1$ is obtained.

After we have $\{\mathbb{E}T_m, 1 \leq m < n\}$, the AFS can be constructed through Equation 2. The detailed pseudocode for implementing the algorithm is listed in Table 1.

Approach 2

For some population histories, the analytical form of the time scaling function $g(t)$ can be achieved, but the inverse function $g^{-1}(t)$ is not tractable. An alternative approach can be applied to obtain $\mathbb{E}T_m$ for such cases through the following procedures. For each $T_m, 1 \leq m < n$, we have the non-linear equation,

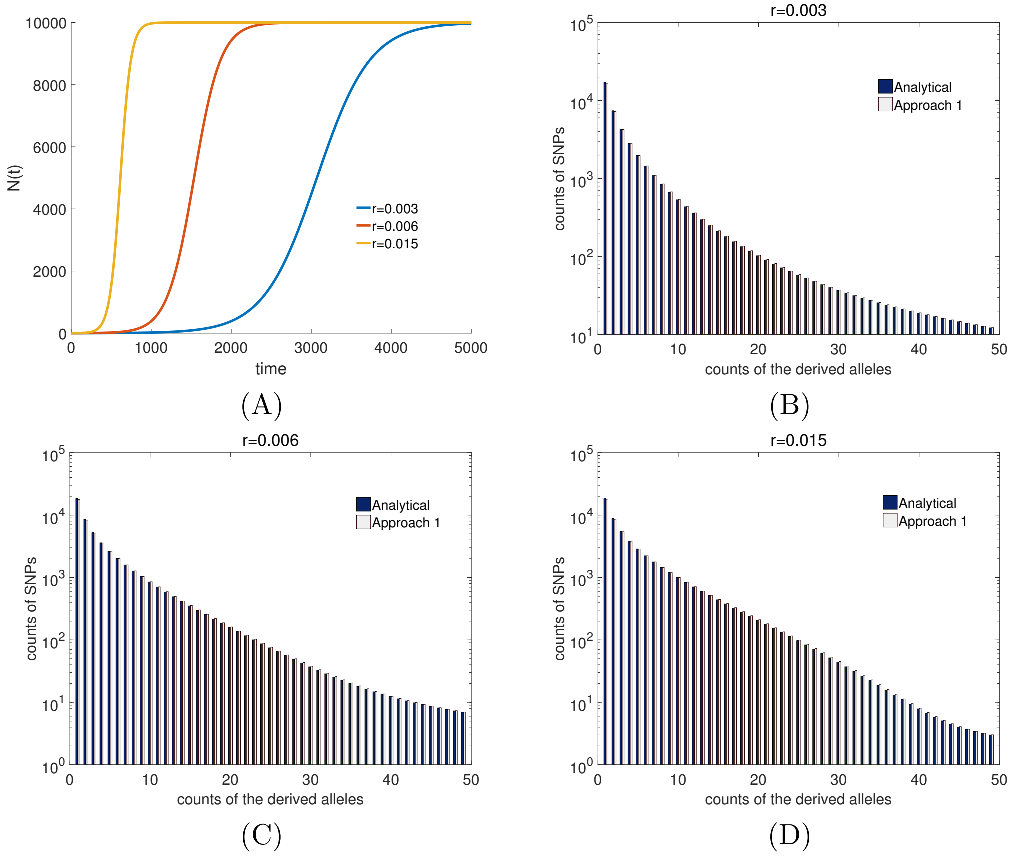
$$g(T_m) - \mu_m = 0, 1 \leq m < n. \tag{12}$$

The above non-linear equations can be solved using numerically algorithms to obtain T_m. such as Newton-Raphson (Press *et al.*, 1992). In this paper, we adopt two approaches implemented in MATLAB. The first one is the fzero function, which implements Dekker's algorithm as a combination of bisection, secant, and inverse quadratic interpolation methods (Brent, 2013). The second is the fminsearch function, which uses the simplex search method of Lagarias *et al.* (1998)

This approach usually takes more time than Approach 1, as for each coalescence time T_m, we need to solve the corresponding equation iteratively. Furthermore, the number of equations and the computational complexity increase with the sample size, and thus Approach 2 is more suitable for small samples.

Results

Various population growth models have been proposed to approximate the ancient population history of humans and other species. For example, Gazave *et al.* (2014) proposed a



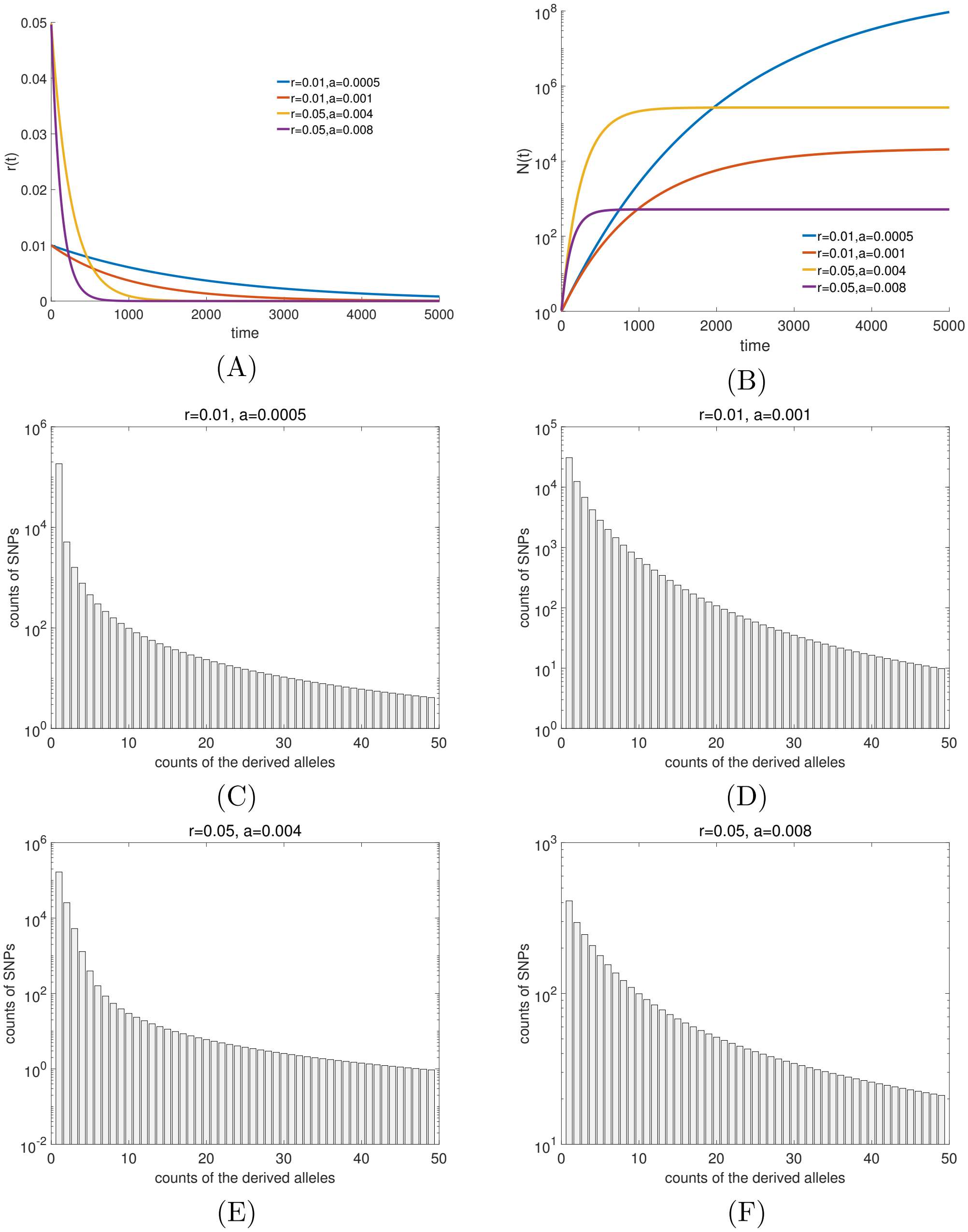
(A)

(B)

(C)

(D)

Figure 1: The population growth rate (A) and population size (B) as a function of time, and the allele frequency spectrum (C) of the logistic growth model for three growth rates: $\gamma = 0.003, 0.006$ and 0.015. The other parameters are: $N_k = 10,000$ and $T = 5000$.



(A)

(B)

(C)

(D)

(E)

(F)

Figure 2: The population growth rate (A) and population size (B) as a function of time, and the allele frequency spectrum (C) of the Gompertz model for four parameter settings: $\gamma = 0.01, \alpha = 0.005; \gamma = 0.01, \alpha = 0.001; \gamma = 0.05, \alpha = 0.004$ and $\gamma = 0.05, \alpha = 0.008$. The other parameters are: $N_0 = 1$, $T = 5000$.

Method	Running time (seconds)		
	Exponential	Logistic	Gompertz
n=10			
analytical calculation	0.000004	0.110637	-
finite sum approximation (approach 1)	0.000084	0.000205	0.000204
fzero (bisection+interpolation)	0.003677	0.004618	-
fminsearch (downhill simplex)	0.019621	0.019983	-
n=50			
analytical calculation	0.000005	0.614442	-
finite sum approximation (approach 1)	0.000087	0.000188	0.000208
fzero (bisection+interpolation)	0.034866	0.020172	-
fminsearch (downhill simplex)	0.063361	0.106250	-
n=100			
analytical calculation	0.000006	1.265550	-
finite sum approximation (approach 1)	0.000087	0.000194	0.000206
fzero (bisection+interpolation)	0.068639	0.041638	-
fminsearch (downhill simplex)	0.126790	0.214588	-
n=500			
analytical calculation	0.000031	7.22106	-
finite sum approximation (approach 1)	0.000145	0.000226	0.000209
fzero (bisection+interpolation)	0.377974	0.231884	-
fminsearch (downhill simplex)	0.737102	1.219030	-

Table 2: Comparison of running times between different methods for three population growth models and four different sample sizes (10, 50, 100 and 500). For the Gompertz model, only the results of the finite-sum approximation are available.

five-scenario model for the European population, including two stages of population bottlenecks and a very recent exponential growth. The simple exponential population growth model may be the most commonly used model. It assumes a constant growth rate, which is valid when space and resources are unlimited. The exponential growth model is a good approximation for the early stage of humans, bacteria, and most populations. In cancer evolution studies, models with more parameters were developed to describe tumor growth (Benzekry *et al.*, 2014). These models are complicated by modifying growth rates with carrying capacity or other factors, e.g., the logistic growth model and Gompertz model.

In this section, we use exponential, logistic and Gompertz growth models to illustrate the usage of our proposed approach. For the exponential growth model, $N(t) = N_0 e^{-rt}$, it is straightforward to analytically derive the expected coalescence times $\mathbb{E}T_m$ (Equation 9). We then compared the running time of the three approaches (including the analytical approach, the finite-sum approximation, and Approach 2) for the model with the two parameters $N_0 = 100,000$ and $r = 0.003$. For Approach 2, we adopted two numerical methods: the bisection + interpolation method implemented in the MATLAB function fzero and the downhill simplex method implemented in the MATLAB function fminsearch. The time was averaged over 1000 repeats running in MATLAB and is recorded in Table 2. The detailed results for the logistic growth and Gompertz model are elaborated below.

Logistic growth

Compared with the exponential growth model, the logistic growth model regulates the growth rate with a factor $(1 - \frac{N(\tilde{t})}{N_k})$, in which N_k is the carrying capacity. It thus has a sigmoid shape and reaches an equilibrium size of N_k instead of unlimited growth (see Figure 1(A)). A logistic growth model is consistent with the population dynamics of many organisms and is widely used in ecological research. Let γ be the maximum population growth rate (aka, intrinsic growth rate); for a population under logistic growth,

$$\frac{dN(\tilde{t})}{d\tilde{t}} = \gamma(\frac{N_k - N(\tilde{t})}{N_k})N(\tilde{t}). \tag{13}$$

Note that in the above equation, time is measured forward (from the past to the present), and we denote it with $\tilde{t}$ to distinguish it from the backward time in other sections. The population size $N(\tilde{t})$ follows a logistic curve,

$$N(\tilde{t}) = \frac{N_k}{(1 - e^{-\gamma\tilde{t}}) + N_k e^{-\gamma\tilde{t}}}. \tag{14}$$

After changing the variable of forward time $\tilde{t}$ to backward time t, we have

$$N(t) = \frac{N_k e^{\gamma T}}{e^{\gamma T} + (N_k - 1)e^{\gamma t}}, \tag{15}$$

and the model includes three free parameters: N_k, γ and T.

Given the population history function $N(t)$, we can derive the time-scaling function for the logistic growth model,

$$\begin{aligned} g(t) &= \int_0^t \frac{1}{N(u)} du \\ &= \frac{e^{-\gamma T}(e^{\gamma t} - 1)(N_k - 1) + \gamma N_k t}{N_k \gamma}, \end{aligned} \tag{16}$$

and we further obtain its inverse function,

$$g^{-1}(\tau) = \frac{-W\left(e^{(N_k-1)e^{-rT}} + N_k\gamma\tau - \gamma T\right) + N_k r\tau + N_k - 1}{r}, \tag{17}$$

where $W(\cdot)$ is the Lambert W function, which is calculated numerically.

According to Chen and Chen (2013), the expected coalescence time $\mathbb{E}T_m = g^{-1}(\mu_m)$ can be obtained from Equation 17. We can also calculate it through Approaches 1 and 2 of the previous section. In Figure 1, we present the AFS generated from Equation 17 ("Analytical calculation") and Approach 1 ("Approach 1") for $N_k = 10,000$, $T = 5000$ and three different growth rates $\gamma = 0.003, 0.006$ and 0.015. We also obtained the AFS using Approach 2 and compared the running time for a specific parameter setting ($N_k = 10,000$, $T = 5000$, and $\gamma = 0.005$) for three approaches (Table 2).

Gompertz growth

The Gompertz model is another widely used model to approximate population dynamics. It was originally proposed to explain human mortality (Gompertz, 1825) and is also used to describe the population growth of other species, including bacteria, animals, and plants (Tjørve and Tjørve, 2017). The Gompertz model was found to fit well the growth of breast cancer and 19 other tumor cell populations (Laird, 1964; Norton *et al.*, 1976; Norton, 1988). One of its forms is

$$\frac{dN(\tilde{t})}{d\tilde{t}} = \gamma(\tilde{t})N(\tilde{t}), \text{ with } \frac{d\gamma}{d\tilde{t}} = -\alpha\gamma(\tilde{t}), \tag{18}$$

where γ_0 is the initial growth rate; N_0 is the initial population size when it started to grow; and α can be viewed as the exponentially decaying rate of the growth rate. The population history $N(\tilde{t})$ is determined by four parameters: N_0, γ_0, α and T, the duration since the population growth began. The solution of the above differential equation is

$$N(\tilde{t}) = N_0 \exp\left(\frac{\gamma_0}{\alpha}(1 - e^{-\alpha\tilde{t}})\right). \tag{19}$$

It is unfeasible to derive the time-scaling function $g(t)$ and its inverse function $g^{-1}(t)$ for the Gompertz model. Therefore, we have no analytical calculation or numerical solution (Approach 2) of the coalescence times for the Gompertz model. In Figure 2 (A) and (B), we show the growth rates and population size trajectories as a function of time for four parameter settings: $\gamma_0 = 0.01, \alpha = 0.0005; \gamma_0 = 0.01, \alpha = 0.001; \gamma = 0.05, \alpha = 0.004$; and $\gamma = 0.05, \alpha = 0.008$. The corresponding AFS for $n = 50$ haplotypes is presented in Figure 2 (C)-(F). The running times of Approach 1 for a specific parameter setting ($T = 5000$, $r = 0.01$, $\alpha = 0.001$ and $N_0 = 1$) and with different sample sizes (10, 50, 100, and 500) are presented in Table 2.

Computing time

We compared the computing times for Approach 1 (finite-sum approximation), Approach 2 and the analytical approach. For Approach 2, we used several methods for solving the non-linear equations, including the bisection+interpolation (implemented in the MATLAB function fzero) and the downhill simplex (implemented in fminsearch) methods. All the comparisons are in MATLAB. We investigated three population growth models: the exponential growth, logistic growth and Gompertz growth model. The running times for constructing the coalescence times $T_m, 1 \leq m < n$, for four sample sizes $n = 10, 50, 100$ and 500 were recorded. The running time was averaged over 1000 repeats, as listed in Table 1 (in seconds).

A trend in Table 2 worth noting is that the finite-sum approximation is very fast. The running time is close to that of the analytical calculation, nearly of the same magnitude, and much faster than that of numerical approaches (Approach 2). The only outlier is the logistic model,for which the finite-sum approximation is much faster than the analytical approach. This is because the analytical form of the $g(t)$ function for the logistic model consists of the Lambert W function, which is calculated numerically and is time-consuming.

Second, note that the running time of the finite-sum approximation approach is nearly constant with increasing sample size n. As we mentioned above, the computational complexity is $O(1)$, and thus, it is insensitive to the sample size. This guarantees the computational efficiency of the approach when the sample size becomes large, enabling its application to large-sample data analysis.

The numerical approach for solving the $g(t)$ function (Approach 2) also works efficiently but is more time-consuming than the finite-sum approximation approach for all three population growth models. Furthermore, the running time increases with the sample size n, as the number of nonlinear equations to solve increases linearly with n.

Conclusion

The allele frequency spectrum is informative for population genetic inference. Various AFS-based methods have been developed for inferring population history and detecting natural selection in the past years. They have gained popularity with the abundance of genomic sequencing data (e.g., Bustamante *et al.* (2001); Griffiths and Tavaré (1998); Polanski and Kimmel (2003); Marth *et al.* (2004); Nielsen *et al.* (2005); Williamson *et al.* (2005); Gutenkunst *et al.* (2009), etc.). Compared with the diffusion-based AFS methods which require approximation of the solutions with numerical approaches, modeling the AFS using coalescent theory is computationally efficient. Most population genetic inference methods using the coalescent likelihood require computationally intensive algorithms for parameter estimation, such as importance sampling or Markov chain Monte Carlo, while the coalescent-based AFS methods only depend on the expected coalescence times, which guarantee the analytical form (Fu, 1995; Griffiths and Tavaré, 1998; Chen, 2012).

The coalescent-based AFS has shortcomings. First, for large samples it is impossible to obtain accurate calculations due to numerical overflow of large coefficients in the hypergeometric series. Second, it is difficult to derive the coalescent-based AFS for complex population histories, which limits its application to simple growth models, such as the exponential growth and n-epoch models. Chen and Chen (2013) showed that for complex demography, we can obtain the expected coalescence times through a linear Taylor expansion approximation, which involves the time-scaling function $g(t)$ and its inverse function $g^{-1}(t)$. The analytical equations of coalescence times derived through this approach are in a simple form and can successfully overcome the numerical issue for large samples. Furthermore, the time-scaling scheme is technically applicable to arbitrary complex population histories. However, in practice, the analytical forms of the population-scaling function $g(t)$ and its inverse function are not achievable for many cases, limiting the applications. For example, in the study of cancer cell growth, various population growth models in complex form were proposed to describe the dynamics of cancer cells, for which the analytical form of AFS is difficult to derive. In this paper, we propose a computational approach, the finite-sum approximation, efficiently solving the problem of Chen and Chen (2013) when the analytical form of the time-scaling function $g(t)$ and its inverse function $g^{-1}(t)$ are not derivable.

We apply the computational approach to three widely used models, including the exponential, logistic and Gompertz growth models to demonstrate its performance. As shown in the Results section, the finite-sum approximation approach is computationally very efficient, and the running time is nearly on the magnitude of that of the analytical approach. Furthermore, the computational time does not increase linearly with the sample size, ensuring its efficiency for AFS of large sample sizes. This is especially attractive for the flexibility to tackle a complex population history that is intractable by using the analytical approach, for example, the Gompertz growth model shown in Table 2. The computational approach presented in this paper is applicable to arbitrary complex population history and significantly enables the application of the coalescent-based AFS approaches to population genetic inference in the genomic sequencing era.

Acknowledgements

I am are grateful to Kun Chen for helpful discussions and to Shilei Zhao for helping with the numerical methods in MATLAB. This project was supported by the National Natural Science Foundation of China (Grant No. 31571370, 91631106 and 91731302), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB13020400) and the "One Hundred Talents Program" of the Chinese Academy of Sciences.

References

Benzekry, S., Lamont, C., Beheshti, A., Tracz, A., Ebos, J. M., Hlatky, L., and Hahnfeldt, P. 2014. Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comp. Biol.*, 10(8): e1003800.

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

Brent, R. P. 2013. *Algorithms for minimization without derivatives*. Courier Corporation.

Bustamante, C., Wakeley, J., Sawyer, S., and Hartl, D. 2001. Directional selection and the site-frequency spectrum. *Genetics*, 159(4): 1779–1788.

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

Chen, H. 2013. Intercoalescence time distribtution of incomplete genealogies in temporally varying populations, and applications in population genetic inference. *Ann. Hum. Genet.*, 77(2): 158–173.

Chen, H. and Chen, K. 2013. Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. *Genetics*, 194(3): 721–736.

Chen, H., Hey, J., and Chen, K. 2015. Inferring very recent population growth rate from population-scale sequencing data: Using a large-sample coalescent estimator. *Molecular Biology and Evolution*, 32(11): 2996–3011.

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., Wheeler, D. A., Sabo, A., Lusk, C., Weiss, K. G., Akbar, H., Cree, A., Hawes, A. C., Newsham, I., Varghese, R. T., Villasana, D., Gross, S., Joshi, V., Santibanez, J., Morgan, M., Chang, K., Hale IV, W., Templeton, A. R., Boerwinkle, E., Gibbs, R., and Sing, C. 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, 1: 131.

Donnelly, P. and Tavaré, S. 1995. Coalescents and genealogical structure under neutrality. *Annual review of genetics*, 29(1): 401–421.

Evans, S., Shvets, Y., and Slatkin, M. 2007. Non-equilibrium theory of the allele frequency spectrum. *Theor. Popul. Biol.*, 71(1): 109–119.

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

Feller, W. 2008. *An introduction to probability theory and its applications*, volume 1. John Wiley & Sons.

Fu, Y. X. 1995. Statistical properties of segregating sites. *Theor. Popul. Biol.*, 48(2): 172–197.

Gao, F. and Keinan, A. 2015. Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. *Genetics*, 202: 235–245.

Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., Boerwinkle, E., Gibbs, R. A., Sing, C. F., Clark, A. G., *et al.* 2014. Neutral genomic regions refine models of recent rapid human population growth. *Proc Natl Acad Sci U S A.*, 111(2): 757–762.

Gompertz, B. 1825. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of contingencies. *Phil Trans Roy Soc*, 115: 513–583.

Gravel, S., Henn, B., Gutenkunst, R., Indap, A., Marth, G., Clark, A., Yu, F., Gibbs, R., Bustamante, C., Altshuler, D., *et al.* 2011. Demographic history and rare allele sharing among human populations. *Proc Natl AcadSci U S A.*, 108(29): 11983–11988.

Griffiths, R. C. 1984. Asymptotic line-of-descent distributions. *Journal of Mathematical Biology*, 21(1): 67–75.

Griffiths, R. C. 2006. Coalescent lineage distributions. *Advances in applied probability*, 38(2): 405–429.

Griffiths, R. C. and Tavaré, S. 1994. Sampling theory for neutral alleles in a varying enviroment. *Philos. Trans. R. Soc. Lond. B*, 344: 403–410.

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

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(10): e1000695.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18: 337–338.

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.

Kimura, M. 1955. Solution of a process of random genetic drift with a continuous model. *Proc Natl Acad Sci U S A.*, 41(3): 144–50.

Kingman, J. 1982. The coalescent. *Stochastic processes and their applications*, 13(3): 235–248.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E. 1998. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1): 112–147.

Laird, A. K. 1964. Dynamics of tumour growth. *British journal of cancer*, 18(3): 490.

Liu, X. and Fu, Y.-X. 2015. Exploring population size changes using snp frequency spectra. *Nat Genet*, 47(5): 555.

Lukić, S., Hey, J., and Chen, K. 2011. Non-equilibrium allele frequency spectra via spectral methods. *Theor. Popul. Biol.*, 79(4): 203–219.

Marth, G., Schuler, G., Yeh, R., Davenport, R., Agarwala, R., Church, D., Wheelan, S., Baker, J., Ward, M., and Kholodov, M. 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc Natl AcadSci U S A.*, 100(1): 376.

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 differeential demographic history in three large world populations. *Genetics*, 2004(166): 351–372.

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

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. D. 2005. Genomic scans for selective sweeps using snp data. *Genome Res.*, 15: 1566–1575.

Nordborg, M. 2001. Coalescent theory pp. 179 - 212 in DJ Balding, M. Bishop, and C. Cannings, eds. Handbook of Statistical Genetics.

Norton, L. 1988. A gompertzian model of human breast cancer growth. *Cancer research*, 48(24 Part 1): 7067–7071.

Norton, L., Simon, R., Brereton, H. D., and Bogden, A. E. 1976. Predicting the course of gompertzian growth. *Nature*, 264(5586): 542.

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(1): 427–36.

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(1): 33–40.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. 1992. Numerical recipes in C: the art of scientific programming. *Section*, 10: 408–412.

Sawyer, S. A. and Hartl, D. L. 1992. Population genetics of polymorphism and divergence. *Genetics*, 132(4): 1161–76.

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(6090): 64–69.

Tjørve, K. M. and Tjørve, E. 2017. The use of gompertz models in growth analyses, and new gompertz-model approach: An addition to the unified-richards family. *PloS One*, 12(6): e0178691.

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L andNielsen, 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.

Wooding, S. and Rogers, A. 2002. The matrix coalescent and an application to human single-nucleotide polymorphisms. *Genetics*, 161(4): 1641–50.

Wooding, S. P., Watkins, W. S., Bamshad, M. J., Dunn, D. M., Weiss, R. B., and Jorde, L. B. 2002. Dna sequence variation in a 3.7-kb noncoding sequence 5' of the CYP1A2 gene: implications for human population history and natural selection. *Am. J. Hum. Genet.*, 71(3): 528–42.

Živković, D. and Stephan, W. 2011. Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theor. Popul. Biol.*, 79(4): 184–191.