Asymptotic Distributions of Coalescence Times and Ancestral Lineage Numbers for Populations with Temporally Varying Size

Hua Chen§ and Kun Chen*

§Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA

*Dana Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA

May 7, 2013
Keywords: coalescent theory, gene genealogy, coalescence time, ancestral lineage, ancestral inference, variable population size.

Corresponding author:

Hua Chen

Department of Epidemiology,
Harvard School of Public Health,
Boston, MA 02115
Phone: (617) 935 2254

Email: hchen@genetics.med.harvard.edu
Abstract

The distributions of coalescence times and ancestral lineage numbers play an essential role in coalescent modeling and ancestral inference. Both exact distributions of coalescence times and ancestral lineage numbers are expressed as the sum of alternating series, and the terms in the series become numerically intractable for large samples. More computationally attractive are their asymptotic distributions, which were derived in GRIFFITHS (1984) for populations with constant size. In this paper, we derive the asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. For a sample of size $n$, denote by $T_m$ the $m^{th}$ coalescent time when $m + 1$ lineages coalesce into $m$ lineages, and $A_n(t)$ the number of ancestral lineages at time $t$ back from the current generation. Similar to the results in Griffiths (1984), the number of ancestral lineages, $A_n(t)$, and the coalescence times, $T_m$, are asymptotically normal, with the mean and variance of these distributions depending on the population size function, $N(t)$. At the very early stage of the coalescent, when $t \to 0$, the number of coalesced lineages $n - A_n(t)$ follows a Poisson distribution, and as $m \to n$, $\frac{n(n-1)T_m}{2N(0)}$ follows a gamma distribution. We demonstrate the accuracy of the asymptotic approximations by comparing to both exact distributions and coalescent simulations. Several applications of the theoretical results are also shown: deriving statistics related to the properties of gene genealogies, such as the time to the most recent common ancestor (TMRCA) and the total branch length (TBL) of the genealogy; and deriving the allele frequency spectrum for large genealogies. With the advent of genomic level sequencing data for large samples, the asymptotic distributions are expected to have wide applications in theoretical and methodological development for population genetic inference.
1 Introduction

Coalescent theory provides a fundamental framework for stochastic modeling and likelihood inference in population genetic studies (Griffiths, 1980; Kingman, 1982a; Hudson, 1990; Nordborg, 2001). A coalescent process can be decomposed into two independent processes, the topology of the gene genealogy and the sequential process of intercoalescence times (Kingman, 1982a). In this paper, we aim to investigate the latter process and two important random quantities associated with this process: the coalescence times and the number of ancestral lineages (Kingman, 1982a). Studying the two quantities is both biologically and theoretically meaningful. First, inferring the coalescence times and the number of ancient lineages of a contemporary sample or population helps to elucidate ancient demographic history, including population admixture, migration and founder effect etc. It can also provide insights into medical studies regarding the origin and genetic architecture of inherited diseases in different populations, as well as to ecological studies, for example, on investigating the process of species invasion (Anderson and Slatkin, 2007; Risch et al., 2003; Dlugosch and Parker, 2007). Second, the distributions of coalescence times and ancestral lineage numbers are the essential components needed to construct a coalescent likelihood, for example, in the allele frequency spectrum-based approaches (Tavaré, 1984; Griffiths and Tavaré, 1998; Polanski and Kimmel, 2003; Chen, 2012).

The exact distribution of the number of ancestral lineages at \( t \) generations ago for \( n \) haplotypes randomly collected at present, \( A_n(t), t \geq 0 \), was derived in Tavaré (1984) under the coalescent for constant populations (Eq. (15) in Section 2; see also Griffiths (1980), Donnelly (1984), Watterson (1984) and Takahata and Nei (1985)). The exact distribution has connections to the Ewens’ sampling formula under the infinitely-many-alleles model (Ewens, 1972). In a later study, the equation was extended to populations with temporally varying size (Griffiths and Tavaré, 1998). The seminal equations in Tavaré (1984) and Griffiths and Tavaré (1998) are very useful in methodology development. However, both exact distributions are expressed as the sums of series with alternating signs, and the coefficients of the series become numerically unstable when \( n > 50 \).

As another important quantity in the coalescent process, the coalescence time, \( T_m \), defined as the time when \( m + 1 \) lineages merge into \( m \) lineages, is well known as a sum of \( n - m \) intercoalescence times. These \( n - m \) intercoalescence times are distributed as independent exponential variables with distinct respective
rates $k(k - 1)/2, k = n, \cdots, m + 1$ under a constant population size model. The analytical expressions of many statistics are derived based on this fact. For populations with time-varying size, the intercoalescence times are no longer independent. GRIFFITHS and TAVARÉ (1998) and POLANSKI et al. (2003) derived the distribution of coalescence times under a temporally variable population size model still as a sum of series, and the evaluation of the coefficients also suffers from the numerical issue when sample size is large.

The numerical problem caused by large sample size becomes an indispensable question with the rapid emergence of large-scale sequencing data for samples of thousands of individuals (MARDIS, 2008; ALT-SHULER et al., 2010; COVENTRY et al., 2010), which on the other hand provides an unprecedented opportunity for population genetic study. Great endeavors are pursued to develop computationally efficient approaches for the analysis of genomic data with large sample size. Most existing coalescent-based inference methods in population genetics rely on sampling approaches with intensive computation, such as importance sampling and Markov Chain Monte Carlo, to integrate over the space of gene genealogies (GRIFFITHS and TAVARÉ, 1994b; FELSENSTEIN et al., 1999), and thus are only applicable for analyzing local genomic regions in small samples. A recently developed method, centered on a coalescent-based joint allele frequency spectrum (JAFS) (CHEN, 2012), gains computational efficiency for the analysis of genomic data from multiple populations, as the author used the derived analytical form of the coalescent-based JAFS instead of the sampling approaches. One of the limitations is that the author derived the JAFS based on TAVARÉ (1984) and GRIFFITHS and TAVARÉ (1998) equations, and the numerical issues of these equations limit the use of the JAFS to small gene genealogies.

GRIFFITHS (2006) simplified the computation of the exact lineage distribution by replacing the sum of alternating series with the hypergeometric function, which has a representation in terms of a complex integral and can be evaluated by numerical integration or simulation. As the distribution is not in simple form, it may intimidate its use for theory and methodology development. POLANSKI and KIMMEL (2003) used the methods of hypergeometric summation to avoid the numerical issue of large $n$ when using the exact distribution of coalescence times to obtain the allele frequency spectrum (AFS) under a time-varying population size model. Their method avoids the calculation of the coefficients in the alternating series that will explode when gene genealogy size increases. However, this approach is designed specifically for calculation of the AFS for some demographic scenarios, and is not a general solution for the numerical instability in
the calculation of the distributions of coalescence times and the number of ancestral lineages. Another way to avoid the calculation of the series with alternating signs is to use the asymptotic approximation instead of the exact distribution. The asymptotic distributions have an additional advantage that they are often in simpler form and are easier for theory establishment.

The asymptotic theories of the coalescence times and the number of ancestral lineages for large gene genealogies in constant populations have been derived by Griffiths (1984). He demonstrated that as \( t \to 0 \) and the sample size \( n \to \infty \), the distributions of \( A_n(t) \) and \( T_m \), converge asymptotically to normal distributions. The essential ingredient in Griffiths’ proof is to apply Lyapunov’s theorem to independently distributed intercoalescence times. For populations with temporally varying size, the validity of Griffiths’ theorems is yet to be addressed, as the intercoalescence times are dependent variables in this case, violating the independence assumption of Lyapunov’s theorem (Billingsley, 2012). However, if we scale the time to account for the fluctuation in population size by \( \int_0^t \frac{ds}{N(s)} \), \( t \geq 0 \), where \( N(\cdot) \) is the population size function over time, the coalescent process on the new time scale is equivalent to the standard coalescent (Kingman, 1982b; Griffiths and Tavaré, 1994b). The theorems for the standard coalescent in Griffiths (1984) can then be borrowed to obtain asymptotic distributions for populations with temporally varying size. Extension of Griffiths’ theorems to populations with time-varying size is very important for population genetic inference, since most ancestral inference is based on the non-equilibrium genetic polymorphism patterns in populations with temporally varying size. Also, the population size and growth rate are themselves demographic parameters of great interest.

In the following sections, we first derive in Section 2 the asymptotic distributions of coalescent times and the number of ancestral lineages for populations with temporally varying size, specifically, for populations under exponential growth. In Section 3 we then compare the asymptotic distributions to exact distributions or coalescent simulations if the exact distributions are difficult to evaluate. We demonstrate that the asymptotic distributions of coalescence times and lineage numbers coincide with both the simulated and exact distributions surprisingly well for a wide range of parameters and for samples with even moderate size. Last, in Section 4, we apply the asymptotic distributions to deriving statistics related to the properties of gene genealogies, such as the expected time to the most recent common ancestor (TMRCA) and the total branch lengths (TBL), and deriving the allele frequency spectrum (AFS) for large samples in simpler
Figure 1: An illustration of the gene genealogy and coalescence times of 5 lineages at present. The coalescence time $T_m$ is defined as the time when $m + 1$ lineages coalesce into $m$ lineages.

analytical form. The paper closes with a discussion in Section 5.

2 Asymptotic distributions for coalescence times and ancestral lineage numbers

2.1 Notations and summary

Consider a sample of $n$ lineages (haplotypes) randomly drawn from the contemporary population. Let $N(t)$ be the deterministic haploid population size at $t$ generations ago. The historical population size $N(t)$ is assumed to be large enough to satisfy the Kingman’s coalescent assumption ($N(t) >> n$). For simplicity of notation, let $N_0 \equiv N(0)$ be the size of the contemporary population. Following Griffiths and Tavaré (1994a), the relative size function $\lambda(t)$ is defined as

$$\lambda(t) = \frac{N(t)}{N_0}. \quad (1)$$

Two random quantities we are investigating are the coalescence times and ancestral lineage numbers. Denote by $T_m, 1 \leq m \leq n$, the coalescence time when $m + 1$ lineages merge into $m$ lineages, with $T_n \equiv 0$ (Fig. 1). It is known that the intercoalescence time, $W_m = T_{m-1} - T_m$, or the time length
of gene genealogies during which there are $m$ lineages, is distributed as an exponential variable with rate \( \frac{m(m-1)}{2N_0} \) for populations with constant size $N_0$ (Fu, 1995). The coalescence time $T_m$ can also be written as $T_m = \sum_{k=m+1}^{n} W_k$.

Denote by $A_n(t)$ the number of ancestral lineages at $t$ generations back in the past. In an ancestral process where both coalescent and mutation events can reduce the number of lineages, $A_n(t)$ is referred to as the number of non-mutant lineages at $t$ generations back from the present (Griffiths, 1984). In this context, we consider the genealogical history of only coalescent events, in which mutations are treated separately and assumed to occur independently following a Poisson process along the branches of a given gene genealogy. The random process $\{A_n(t), t \geq 0\}$ is a pure-death process that jumps from state $m$ to $m-1$ with rate $m(m-1)/2N_0$, $2 \leq m \leq n$ (Kingman, 1982a). All the random variables, $T_m$, $W_m$ and $A_n(t)$, are defined in a coalescent process with time in units of generations. Let $\tau = g(t) = \int_0^t \frac{1}{N(u)} \, du$ be the time scaled at rate $N(t)$, specifically for populations with constant size, $\tau = \frac{t}{N_0}$. We use the notations $\hat{T}_m$, $\hat{W}_m$ and $\hat{A}_n(\tau)$ to denote the coalescence time, intercoalescence time and number of ancestral lineages in the standard coalescent process with time scaled by the constant population size $N_0$, which is also referred to as Kingman’s $n$-coalescent process in the context.

In the following, we aim to develop the asymptotic distributions of coalescence times, $T_m$, and ancestral lineage numbers, $A_n(t)$, for populations with temporally varying size $N(t)$. The main results include:

1. for a sample of size $n$, as $n \to \infty$, $n/m \to a$, $1 < a < \infty$, the $m^{th}$ coalescence time, $T_m$, is asymptotically normal, with the mean and variance depending on the historical population size (Eqs. (8) and (9));

2. as $g(t) \to 0$, $n \to \infty$, $\frac{1}{2}ng(t) \to \alpha$, $0 < \alpha \leq \infty$, the number of ancestral lineages at time $t$, $A_n(t)$, is asymptotically normal, with the mean and variance provided in Eqs. (19) and (20);

3. at the very early stage of the coalescent, when $t \to 0$ more rapidly, $n \to \infty$, the number of coalesced lineages $n - A_n(t)$ follows a Poisson distribution, and when $m \to n$, $n - m$ is bounded above, $\frac{n(n-1)T_m}{2N_0}$ follows a gamma distribution.
2.2 Asymptotics of coalescence times

Under a constant population size model, the intercoalescence times, $\hat{W}_m, 2 \leq m \leq n$, are independent exponential variables with respective rates $\left(\frac{m}{2}\right)$. Griffiths (1984) proved the asymptotic distribution of coalescence times by applying Lyapunov’s version of the central limit theorem to the sum of independent intercoalescence times. Denote by $\hat{T}_m$ the $m^{th}$ coalescence time in a standard $n$-coalescent process, scaled by population size $N_0$ (Ewens, 2004). By Theorem 1 of Griffiths (1984), under the conditions $n \to \infty$, $m \to \infty$, while $n/m \to a$, $1 < a \leq \infty$, $\hat{T}_m$ is asymptotically normal, and the mean and the variance of $\hat{T}_m$ are

$$\mu_m = \sum_{k=m+1}^{n} \frac{2}{k(k-1)} = 2\left(\frac{1}{m} - \frac{1}{n}\right),$$

and

$$\sigma_m^2 = \sum_{k=m+1}^{n} \frac{4}{k^2(k-1)^2} = 4\sum_{k=m+1}^{n} \left[\frac{1}{(k-1)^2} + \frac{1}{k^2} - \frac{2}{k(k-1)}\right] = 4\left\{\psi_1(m) - \psi_1(n) + \psi_1(m+1) - \psi_1(n+1) - 2(m^{-1} - n^{-1})\right\},$$

where the trigamma function $\psi_1(z) = \frac{d^2}{dz^2} \ln \Gamma(z)$ and $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1} \, dt$. Note that the result above is a little different from what was originally shown in Griffiths (1984) in that the effect of mutations on lines of descent is not considered in Eqs. (2) and (3). We take the strategy of constructing the genealogy first, and then given the branch lengths of the genealogy, modeling mutations as a Poisson process with the rate proportional to the specific branch length.

Under a variable population size model, the intercoalescence times are no longer mutually independent. We assume that the population evolves according to the Wright-Fisher model, and its size changes over time deterministically, that is, the population size is different but known at each generation. The joint distribution of coalescence times $(T_m, \ldots, T_{n-1})$ for populations with temporally varying size is (Griffiths and Tavaré, 1998):

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

The marginal p.d.f. of coalescence times $f_{T_m}$ was derived explicitly by Polanski et al. (2003) through
expanding an integral transform of the marginal p.d.f. into partial fractions. An equivalent equation in different form can be derived based on the definition of the $n$-coalescent and a pure-death process (see Griffiths (2006) and Appendix A of Chen (2012) for details of derivation):

$$f_{T_m}(t) = \mathbb{P}(A_n(t) = m + 1) \frac{(m + 1)m}{2N(t)}.$$  

(5)

As $\mathbb{P}(A_n(t) = m + 1)$ was derived from the expansion of the transition function (see Eqs. (15) and (16)), which involves the sum of alternating series and is numerically unstable for large $n$, the exact distribution may be practically difficult to use in the ancestral inference for large samples. In the following, we aim to derive the asymptotic distribution of coalescence times for temporally varying populations.

It is known in the coalescent literature that the coalescence time rescaled at rate $\frac{1}{N(t)}$, $g(T_m) = \int_0^{T_m} \frac{1}{N(u)} du$, follows the distribution of coalescence time in the standard Kingman’s $n$-coalescent, and the scaled intercoalescence times,

$$g(T_{m-1}) - g(T_m) = \int_{T_m}^{T_{m-1}} \frac{1}{N(u)} du, 2 \leq m \leq n,$$

are mutually independent exponential variables with the rate of $\binom{m}{2}$. The coalescent process under a variable population size model, as sample size $n \to \infty$, is still a Kingman’s coalescent since we assume the population size tends to infinity more quickly than the sample size, in other words, $n/N(t) \to 0$, which makes the condition of Kingman’s coalescent $n \ll N(t)$ still satisfied. More detailed discussions on time scaling in a variable population size coalescent model can be found in Kingman (1982b), Griffiths and Tavaré (1994a), Donnelly and Tavaré (1995), and Nordborg (2001).

We start with a Taylor expansion of $g(T_m)$ at $g^{-1}(\mu_m)$,

$$g(T_m) = \mu_m + g'(g^{-1}(\mu_m))(T_m - g^{-1}(\mu_m)) + \frac{g''(g^{-1}(\mu_m))}{2}(T_m - g^{-1}(\mu_m))^2 + O((T_m - g^{-1}(\mu_m))^3),$$

(6)

where $g^{-1}(\cdot)$, $g'(\cdot)$ and $g''(\cdot)$ represent the inverse function, the first derivative and the second derivative of function $g$ respectively. The remainder term $\frac{g''(g^{-1}(\mu_m))}{2}(T_m - g^{-1}(\mu_m))^2 + O((T_m - g^{-1}(\mu_m))^3)$, or $O((T_m - g^{-1}(\mu_m))^2)$ in Eq. (6) is ignorable as $n/m \to a, n \to \infty$, because $g(T_m)$ follows the same
distribution as \( \hat{T}_m \) and \( \hat{T}_m \to \mu_m \) by the asymptotic properties of \( \hat{T}_m \) shown in Theorem 1 of Griffiths (1984).

Next, by Eq. (6) and ignoring the remainder term, we have

\[
\frac{T_m - g^{-1}(\mu_m)}{(\sigma_m / g'(g^{-1}(\mu_m)))} \to \frac{g(T_m) - \mu_m}{\sigma_m}.
\] (7)

As \( \frac{g(T_m) - \mu_m}{\sigma_m} \to N(0, 1) \), the limiting distribution of \( T_m \) can then be approximated by a normal distribution with the mean

\[
E(T_m) = g^{-1}(\mu_m),
\] (8)

and variance

\[
\text{Var}(T_m) = \frac{\sigma_m^2}{(g'(g^{-1}(\mu_m)))^2}.
\] (9)

Substituting Eqs. (2) and (3) into Eqs. (8) and (9) yields the mean and variance for the asymptotic distribution of coalescence times, \( T_m, 1 \leq m \leq n - 1 \).

When the population is under exponential growth with rate \( \gamma \), that is, \( N(t) = N_0 e^{-\gamma t} \), it is straightforward to write the scaling function as

\[
g(t) = \frac{e^{\gamma t} - 1}{N_0 \gamma}.
\] (10)

The inverse and first derivative function of \( g \) are \( g^{-1}(\tau) = \frac{\ln(N_0 \gamma \tau + 1)}{\gamma} \) and \( g'(t) = \frac{e^{\gamma t}}{N_0 \gamma} \) respectively. Using Eqs. (8) and (9), we have the mean

\[
E(T_m) = g^{-1}(\mu_m) = \frac{\ln(N_0 \gamma \mu_m + 1)}{\gamma} = \frac{1}{\gamma} \ln(2N_0 \gamma (m^{-1} - n^{-1}) + 1),
\] (11)
and variance

\[
\text{Var}(T_m) = \frac{\sigma_m^2}{(g'(g^{-1}(\mu_m)))^2} = \exp\left\{2\gamma \frac{1}{\gamma} \ln(2N_0 \gamma (m^{-1} - n^{-1}) + 1)\right\}/N_0^2 = \left(\frac{N_0}{2N_0 \gamma (m^{-1} - n^{-1}) + 1}\right)^2 \sigma_m^2 = \frac{4N_0^2 \left(\psi_1(m) - \psi_1(n) + \psi_1(m + 1) - \psi_1(n + 1) - 2(m^{-1} - n^{-1})\right)}{(2N_0 \gamma (m^{-1} - n^{-1}) + 1)^2}.
\] (12)

**Bias of the asymptotic mean of** $T_m$: Since the linear approximation is used in the above proof, there exists a bias between the derived asymptotic mean and the true mean of $T_m$. Here we quantify the magnitude of the bias specifically for the exponential growth model as an example. Using a Taylor expansion,

\[
T_m = g^{-1}(g(T_m)) = g^{-1}(\mu_m) + (g^{-1})'(\mu_m)(g(T_m) - \mu_m) + \frac{(g^{-1})''(\mu_m)}{2}(g(T_m) - \mu_m)^2 + O((g(T_m) - \mu_m)^3),
\]

and taking expectation at both sides, it can be seen that

\[
\mathbb{E}(T_m) - g^{-1}(\mu_m) = \frac{(g^{-1})''(\mu_m)}{2} \mathbb{E}(g(T_m) - \mu_m)^2 + o(1) = -\frac{N_0^2 \gamma \sigma_m^2}{2(N_0 \gamma \mu_m + 1)^2} + o(1) = -\frac{\sigma_m^2}{2\gamma \mu_m^2} + o(1).
\] (13)

By Theorem 1 in Griffiths (1984), $\mu_m$ is on order $m^{-1}$ and $\sigma_m^2$ is on order $m^{-3}$. Therefore, the bias of the asymptotic mean is on order $m^{-1}$, which shrinks to zero as $n/m \to a, n \to \infty$.

2.3 Asymptotics of ancestral lineage numbers

Tavaré (1984) derived the exact distribution of the number of ancestral lineages at time $t$ in the past for the coalescent with constant population size, by using a spectral expansion of the transition function that is associated with the death process $\{A_n(t), t \geq 0\}$ (see also Griffiths (1980); Donnelly (1984);
\begin{equation}
\mathbb{P}(A_n(t) = m) = \sum_{i=m}^{n} \frac{(-1)^{i-m}(2i-1)m(i-1)n[i]}{i!(i-m)!n[i]} e^{-i(i-1)t/2N}, \quad 0 < m \leq n,
\end{equation}

where \( n(i) = n(n+1) \ldots (n+i-1), \) \( i \geq 1; \) \( n[0] = 1, \) and \( n[i] = n(n-1) \ldots (n-i+1), \) \( i \geq 1; \) \( n[0] = 1 \) are the rising and falling factorial functions. Griffth's and Tavaré (1998) further generalized Eq. (15) to populations with variable size. Then for populations with temporally varying size, the distribution of the number of ancestral lineages becomes (Griffith's and Tavaré, 1998):

\begin{equation}
\mathbb{P}(A_n(t) = m) = \sum_{i=m}^{n} \frac{(-1)^{i-m}(2i-1)m(i-1)n[i]}{i!(i-m)!n[i]} e^{-i(i-1)t/2N} \int_0^t \lambda(u) du.
\end{equation}

In addition to the coalescence times, Griffith's (1984) investigated the asymptotics of ancestral lineage numbers for populations with constant size. If omitting mutation for the same reason as in Section 2.2, by Griffith's Theorem 2, \( \dot{A}_n(\tau) \) has an asymptotic normal distribution as \( \tau \to 0, n \to \infty, \) \( \frac{1}{2}n\tau \to \alpha, 0 < \alpha < \infty, \) with a mean of

\begin{equation}
\mu(\tau) = \frac{2\eta}{\tau},
\end{equation}

and variance

\begin{equation}
\sigma^2(\tau) = 2\eta\tau^{-1}(\eta + \beta)^2\{1 + \eta/(\eta + \beta) - \eta/\alpha - \eta/(\alpha + \beta) - 2\eta\}\beta^{-2},
\end{equation}

where \( \beta = -\frac{1}{2}\tau, \eta = \alpha\beta/\{\alpha(e^\beta - 1) + \beta e^\beta\}. \) In the following, we extend Griffith's conclusion of the asymptotic distribution for ancestral lineage numbers to populations with temporally varying size.

We observe that \( \mathbb{P}(A_n(t) \leq m) = \mathbb{P}(T_m \leq t). \) As \( g \) is a monotone continuous function, we have \( \mathbb{P}(T_m \leq t) = \mathbb{P}(g(T_m) \leq g(t)) = \mathbb{P}(\dot{A}_n(\tau) \leq m), \) where \( \tau \equiv g(t) \) as defined in last section. By Griffith's Theorem 2, as \( n \to \infty, \tau \to 0, \) and \( \frac{1}{2}n\tau \to \alpha, \) the ancient lineage number of the coalescent process \( \dot{A}_n(\tau) \to d N(u(\tau), \sigma^2(\tau)), \) where \( \mu(\tau) \) and \( \sigma^2(\tau) \) are given in Eqs. (17) and (18). Therefore, \( \mathbb{P}(A_n(t) \leq m) \to \mathbb{P}(Z \leq \frac{m-u(\tau)}{\sigma(\tau)}), \) where \( Z \) is a standard normal.

The mean and variance of the limiting distribution of \( A_n(t) \) can be obtained by mapping back to the
original time scale, as follows,

$$\mathbb{E}(A_n(t)) \to u(\tau) = u(g(t)) = 2\eta(g(t))^{-1},$$ \hspace{1cm} (19)

and

$$\text{Var}(A_n(t)) \to \sigma^2(\tau) = \sigma^2(g(t)) = 2\eta(g(t))^{-1}(\eta + \beta)^2\{1 + \eta/(\eta + \beta) - \eta/\alpha - \eta/(\alpha + \beta) - 2\eta\}\beta^{-2},$$ \hspace{1cm} (20)

where $\alpha = \lim_{n \to \infty, t \to 0} \frac{1}{2} n g(t), \beta = -\frac{1}{2} g(t), \text{ and } \eta = \frac{\alpha \beta}{\{\alpha(e^\beta - 1) + \beta e^\beta\}}$.

When the population is under exponential growth with rate $\gamma$ and scaling function as in Eq. (10), plugging Eq. (10) into Eqs. (19) and (20), $A_n(t)$ then has an asymptotic mean

$$u(g(t)) = \frac{2\eta N_0 \gamma}{e^{\gamma t} - 1},$$ \hspace{1cm} (21)

and variance

$$\sigma^2(g(t)) = \frac{2\eta N_0 \gamma}{e^{\gamma t} - 1}(\eta + \beta)^2\{1 + \eta/(\eta + \beta) - \eta/\alpha - \eta/(\alpha + \beta) - 2\eta\}\beta^{-2},$$ \hspace{1cm} (22)

where $\alpha = \lim_{n \to \infty, t \to 0} \frac{1}{2} \frac{n(e^\gamma - 1)}{N_0 \gamma}, \beta = -\frac{1}{2} \frac{e^\gamma - 1}{N_0 \gamma}, \text{ and } \eta = \frac{\alpha \beta}{\{\alpha(e^\gamma - 1) + \beta e^\gamma\}}$.

As no linear approximation is used to derive the asymptotic distribution of $A_n(t)$, if $u(\tau)$ and $\sigma^2(\tau)$ were the exact mean and variance of $\dot{A}_n(\tau), u(g(t))$ and $\sigma^2(g(t))$ should be the exact mean and variance of $A_n(t)$. Here we use the asymptotic mean and variance in Griffiths (1984) for $u(\tau)$ and $\sigma^2(\tau)$, and this results in the bias in our derived asymptotic mean and variance of $A_n(t)$.

2.4 Asymptotics of coalescence times and ancestral lineage numbers at the early stage of the coalescent

At the early stage of the coalescent, or $m \to n$ and $t \to 0$, the above normal distributions may not well approximate the exact distributions of coalescence times and ancestral lineage numbers. We derive their asymptotics in this section.
As the scaled coalescence time \( g(T_m) \) follows the distribution of coalescence times in a standard coalescent process, \( g(T_m) - g(T_{m+1}) \sim \text{Exponential}(\binom{m+1}{2}) \). Multiplying \( g(T_m) - g(T_{m+1}) \) by \( \frac{n(n-1)}{2} \),

\[
\frac{n(n-1)}{2} (g(T_m) - g(T_{m+1})) \sim \text{Exponential}(\frac{m+1}{2}) \rightarrow \text{Exponential}(1) \text{ as } m \rightarrow n \text{ and } n \rightarrow \infty.
\]

By the Mean-Value Theorem, \( g(T_m) - g(T_{m+1}) = (T_m - T_{m+1})g'(\xi_{m+1}) \), where \( T_{m+1} \leq \xi_{m+1} \leq T_m \). As \( m \rightarrow n \) and \( n \rightarrow \infty \), we have \( T_m \rightarrow 0 \) and \( g'(\xi_{m+1}) = \frac{1}{N(\xi_{m+1})} \rightarrow \frac{1}{N_0} \). Subsequently,

\[
\frac{n(n-1)}{2} \sum_{k=m+1}^{n} (g(T_{k-1}) - g(T_k)) = \frac{n(n-1)}{2} \sum_{k=m+1}^{n} (T_{k-1} - T_k) g'(\xi_k).
\]

Taking limits \( n \rightarrow \infty \), \( m \rightarrow n \) at both sides of Eq. (23), we obtain \( \frac{n(n-1)}{2N_0} T_m \rightarrow \text{Gamma}(n-m, 1) \).

Next, we derive the asymptotic distribution for the number of coalesced lineages, \( n - A_n(t) \), for a population with time-varying size \( N(t) \). Given that \( \mathbb{P}(n - A_n(t) \geq n - j) = \mathbb{P}(n - \hat{A}_n(g(t)) \geq n - j) = \mathbb{P}(n - \hat{A}_n(\tau) \geq n - j) \), and Theorem 6 in Griffiths (1984), \( n - A_n(t) \) asymptotically follows the Poisson distribution with mean \( \nu = \frac{1}{2} n(n-1)g(t) = \frac{1}{2} n(n-1) \int_0^t \frac{1}{N(u)} du \).

### 3 Numerical results

The accuracy of the asymptotic distributions of \( T_m \) and \( A_n(t) \) derived in Section 2 is examined by comparing their distributional properties to both exact distributions and coalescent simulations. If the analytical formulas of the exact distributions for \( T_m \) and \( A_n(t) \) are available and can be computed without numerical issues, for example, the mean and variance of the ancestral lineage number distribution, we use the analytical results instead of coalescent simulations in the comparison. Otherwise, we seek to compare the asymptotic distributions with simulated distributions. The coalescent simulator ms is modified to output coalescence times and the number of ancestral lineages for each simulated gene genealogy (HUDSON, 2002). In the simulation, the contemporary population size \( N_0 \) is chosen to be \( 2 \times 10^6 \), and a wide range of exponential growth rate \( \gamma \), time \( t \) and sample size \( n \) is investigated.

#### 3.1 Coalescence times

We first examine the numerical accuracy of the asymptotic probability density function (p.d.f.) of coalescence times as provided in Section 2.2. We show the asymptotic p.d.f.'s of several coalescence times for
In Tables 1 and S1, a more thorough simulation study is presented. The simulation is carried out for three different growth rates: $\gamma = 0.001, 0.005$ and 0.01, and five sample sizes: $n = 50, 100, 200, 500$ and 800. For each combination of parameter values, several coalescence times are examined to cover as much of the time span as possible. For example, for $\gamma = 0.001, n = 200$, we examine $T_5, T_20, T_50, T_{100}, T_{150}$ and $T_{195}$. For most of the $T_m$’s examined here, the asymptotic mean and variance approximate those of the simulated distributions very well.

Note that in Tables 1 and S1 the asymptotic mean of $T_m$ is always larger than the simulated mean. This is consistent with our quantified bias in Eq. (14). It can also be observed that both the bias and the relative bias (the bias divided by the simulated mean) of the asymptotic mean of $T_m$ are bigger when $m$ is close to $n$. This can be explained by the inflated second derivative of the scaling function $g^{-1}(\tau)$ evaluated at $\mu_m$ close to 0 (approximately $-\frac{1}{\gamma \mu_m}$, see also Eq. (14)) appearing in the bias term. The detailed derivation of the quantified bias can be seen in Section 2.2.

Another trend worth noting in Tables 1 and S1 are that the relative bias of the mean decreases with increasing sample size $n$. For example, for $\gamma = 0.001$ and the intercoalescence time $T_{50}$, the relative bias is 0.353%, 0.150%, 0.049%, 0.055% and 0.025% for $n = 100, 200, 500, 800$ and 5000 respectively (the last data point not shown in Tables 1 and S1). For a coalescence time $T_m$ with a smaller $m$, the relative bias is
Figure 2: The asymptotic probability density functions of coalescence times in the history for two parameter settings.

(A) and (B) The asymptotic probability density functions of coalescence times in a sample of size 500, collected from a population with $N_0 = 2 \times 10^6$. The growth rate is $\gamma = 0.001$ in (A) and $\gamma = 0.005$ in (B). The spikes from left to right correspond to the p.d.f.’s of $T_{2}, T_{10}, T_{25}, T_{50}, T_{100}, T_{200}, T_{300}$ and $T_{400}$.

(C) and (D) Comparison of the asymptotic probability density function and simulated distribution of coalescence time $T_{10}$ for growth rate being 0.001 in (C) and 0.005 in (D). The histograms were generated by 500 coalescent simulations.

(E) and (F) Comparison of the asymptotic probability density function and simulated distribution of coalescence time $T_{100}$ for growth rate being 0.001 in (E) and 0.005 in (F). The histograms were generated by 500 coalescent simulations.
reduced more slowly. In Tables 1 and S1, the relative bias for $T_5$ does not have obvious trend of decrease for sample sizes up to 800 (for $\gamma = 0.001$, $T_5$ decreases from 0.642% for $n = 50$ to 0.513% for $n = 800$). When the sample size is increased to 5000, the relative bias becomes 0.405% (data not shown in Tables 1 and S1). Although the convergence rates are different for the above two $T_m$'s, the bias of both $T_m$'s shrinks towards zero as $n \to \infty$.

Table 1: Comparison of the asymptotic approximation and simulated results for the mean and standard deviation of the coalescence time $T_m$ ($N_0 = 2.0 \times 10^6$).

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$m$</th>
<th>Sample size n=50</th>
<th></th>
<th>Sample size n=200</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean ($T_m$)</td>
<td></td>
<td>Standard Deviation ($T_m$)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simulation</td>
<td>Asymptotic</td>
<td>Bias (%)</td>
<td>Simulation</td>
</tr>
<tr>
<td>0.001</td>
<td>5</td>
<td>6538.619</td>
<td>6580.639</td>
<td>42.020 (0.643%)</td>
<td>278.445</td>
</tr>
<tr>
<td>0.001</td>
<td>10</td>
<td>5746.244</td>
<td>5771.441</td>
<td>25.197 (0.438%)</td>
<td>223.672</td>
</tr>
<tr>
<td>0.001</td>
<td>20</td>
<td>4775.353</td>
<td>4795.791</td>
<td>20.438 (0.428%)</td>
<td>209.097</td>
</tr>
<tr>
<td>0.001</td>
<td>45</td>
<td>2210.726</td>
<td>2291.412</td>
<td>80.685 (3.650%)</td>
<td>402.151</td>
</tr>
<tr>
<td>0.005</td>
<td>5</td>
<td>1629.978</td>
<td>1637.793</td>
<td>7.816 (0.480%)</td>
<td>56.458</td>
</tr>
<tr>
<td>0.005</td>
<td>10</td>
<td>1470.136</td>
<td>1475.677</td>
<td>5.541 (0.377%)</td>
<td>44.957</td>
</tr>
<tr>
<td>0.005</td>
<td>20</td>
<td>1274.626</td>
<td>1279.719</td>
<td>5.093 (0.400%)</td>
<td>41.512</td>
</tr>
<tr>
<td>0.005</td>
<td>45</td>
<td>742.147</td>
<td>763.298</td>
<td>21.151 (2.850%)</td>
<td>90.637</td>
</tr>
<tr>
<td>0.010</td>
<td>5</td>
<td>884.172</td>
<td>888.198</td>
<td>4.026 (0.455%)</td>
<td>28.072</td>
</tr>
<tr>
<td>0.010</td>
<td>10</td>
<td>804.606</td>
<td>807.122</td>
<td>2.515 (0.313%)</td>
<td>22.611</td>
</tr>
<tr>
<td>0.010</td>
<td>20</td>
<td>706.889</td>
<td>709.091</td>
<td>2.202 (0.312%)</td>
<td>20.902</td>
</tr>
<tr>
<td>0.010</td>
<td>45</td>
<td>439.281</td>
<td>449.857</td>
<td>10.577 (2.408%)</td>
<td>46.789</td>
</tr>
</tbody>
</table>

Continued on next page
### Table 1 – continued from previous page

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$m$</th>
<th>Mean ($T_m$) Simulation</th>
<th>Asymptotic</th>
<th>Bias (%)</th>
<th>Standard Deviation ($T_m$) Simulation</th>
<th>Asymptotic</th>
<th>Bias (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.010</td>
<td>100</td>
<td>529.684</td>
<td>530.330</td>
<td>0.646</td>
<td>0.122%</td>
<td>10.860</td>
<td>10.747</td>
</tr>
<tr>
<td>0.010</td>
<td>150</td>
<td>420.709</td>
<td>421.459</td>
<td>0.751</td>
<td>0.178%</td>
<td>14.020</td>
<td>14.125</td>
</tr>
<tr>
<td>0.010</td>
<td>195</td>
<td>174.898</td>
<td>181.290</td>
<td>6.392</td>
<td>3.655%</td>
<td>37.343</td>
<td>37.428</td>
</tr>
<tr>
<td>0.001</td>
<td>5</td>
<td>6645.512</td>
<td>6679.599</td>
<td>34.087</td>
<td>0.513%</td>
<td>247.504</td>
<td>258.485</td>
</tr>
<tr>
<td>0.001</td>
<td>10</td>
<td>5958.249</td>
<td>5981.414</td>
<td>23.165</td>
<td>0.389%</td>
<td>181.338</td>
<td>184.235</td>
</tr>
<tr>
<td>0.001</td>
<td>50</td>
<td>4328.359</td>
<td>4330.733</td>
<td>2.374</td>
<td>0.055%</td>
<td>85.882</td>
<td>85.933</td>
</tr>
<tr>
<td>0.001</td>
<td>200</td>
<td>2771.704</td>
<td>2772.589</td>
<td>0.885</td>
<td>0.032%</td>
<td>50.952</td>
<td>50.631</td>
</tr>
<tr>
<td>0.001</td>
<td>400</td>
<td>1789.988</td>
<td>1791.759</td>
<td>1.771</td>
<td>0.099%</td>
<td>37.343</td>
<td>37.428</td>
</tr>
<tr>
<td>0.005</td>
<td>5</td>
<td>1651.462</td>
<td>1657.606</td>
<td>6.144</td>
<td>0.372%</td>
<td>50.473</td>
<td>51.749</td>
</tr>
<tr>
<td>0.005</td>
<td>10</td>
<td>1515.254</td>
<td>1517.766</td>
<td>2.512</td>
<td>0.166%</td>
<td>36.394</td>
<td>36.922</td>
</tr>
<tr>
<td>0.005</td>
<td>50</td>
<td>1185.082</td>
<td>1185.918</td>
<td>0.835</td>
<td>0.070%</td>
<td>17.362</td>
<td>17.369</td>
</tr>
<tr>
<td>0.005</td>
<td>200</td>
<td>865.872</td>
<td>866.147</td>
<td>0.275</td>
<td>0.032%</td>
<td>10.690</td>
<td>10.659</td>
</tr>
<tr>
<td>0.005</td>
<td>400</td>
<td>651.296</td>
<td>651.619</td>
<td>0.323</td>
<td>0.050%</td>
<td>10.458</td>
<td>10.386</td>
</tr>
<tr>
<td>0.005</td>
<td>795</td>
<td>28.726</td>
<td>29.206</td>
<td>0.481</td>
<td>1.673%</td>
<td>12.000</td>
<td>12.153</td>
</tr>
<tr>
<td>0.010</td>
<td>5</td>
<td>894.511</td>
<td>898.105</td>
<td>3.594</td>
<td>0.402%</td>
<td>25.293</td>
<td>25.878</td>
</tr>
<tr>
<td>0.010</td>
<td>10</td>
<td>825.938</td>
<td>828.172</td>
<td>2.234</td>
<td>0.270%</td>
<td>18.054</td>
<td>18.465</td>
</tr>
<tr>
<td>0.010</td>
<td>50</td>
<td>661.756</td>
<td>662.141</td>
<td>0.385</td>
<td>0.058%</td>
<td>8.640</td>
<td>8.696</td>
</tr>
<tr>
<td>0.010</td>
<td>200</td>
<td>501.546</td>
<td>501.728</td>
<td>0.182</td>
<td>0.036%</td>
<td>5.500</td>
<td>5.365</td>
</tr>
<tr>
<td>0.010</td>
<td>400</td>
<td>393.125</td>
<td>393.183</td>
<td>0.058</td>
<td>0.015%</td>
<td>5.358</td>
<td>5.295</td>
</tr>
<tr>
<td>0.010</td>
<td>795</td>
<td>26.667</td>
<td>27.343</td>
<td>0.676</td>
<td>2.534%</td>
<td>10.423</td>
<td>10.699</td>
</tr>
</tbody>
</table>

#### 3.2 Number of ancestral lineages

In this subsection, we aim to evaluate how the asymptotic distribution of $A_n(t)$ performs as an approximation to the true distribution. The exact formulas of the first two moments of the ancestral lineage number distribution under a varying population size model were derived using the probability generating function in Tavaré (1984):

$$
\mathbb{E}^{\text{exact}}(A_n(t)) = \sum_{l=1}^{n} (2l-1) \frac{n[l]}{n(l)} \exp \left\{ \frac{-l(l-1)}{2N_0} \int_0^t \frac{1}{\lambda(u)} du \right\},
$$

and

$$
\mathbb{E}^{\text{exact}}(A_n^2(t)) = \sum_{l=1}^{n} (2l-1) \frac{l(l-1) + 1}{n(l)} \frac{n[l]}{n(l)} \exp \left\{ \frac{-l(l-1)}{2N_0} \int_0^t \frac{1}{\lambda(u)} du \right\}.
$$
Unlike the entire exact distribution of the ancestral lineage number, the exact mean and variance of ancestral lineage numbers can be accurately calculated from Eqs. (24) and (25) even for quite large samples, and are assumed to be the gold standard in the comparison. Three different exponential growth rates, $\gamma = 0.001, 0.003$ and 0.01, and five sample sizes, $n = 50, 100, 200, 400$ and 800 are considered here. For each combination of parameter values, the number of lineages are collected at four time points ($t = 100, 500, 1000$ and 2000 for $\gamma = 0.001$ and 0.003; $t = 100, 300, 600$ and 800 for $\gamma = 0.01$).

Tables 2 and S2 present the mean and standard deviation of the number of lineages calculated from the exact formulas and from the asymptotic results for the above chosen parameter values, as well as the bias and relative bias (the bias divided by the exact mean or standard deviation) of the proposed asymptotic mean and standard deviation. As we can see from Tables 2 and S2, the mean and variance obtained from the proposed asymptotic distribution are close to the exact mean and variance for a wide range of time points, growth rates and sample sizes. Note that the asymptotic results are accurate even for a relatively small sample size, e.g., $n = 50$. The robustness of the asymptotic results for various demographic parameters and small sample size assures the application of the asymptotic distribution in theory and methodology development for population genetic inference.

Table 2: Comparison of the asymptotic approximation and exact results for the mean and standard deviation of $A_n(t)$ (population size $N_0 = 2.0 \times 10^6$).

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$t$</th>
<th>Exact Mean ($A_n(t)$)</th>
<th>Asymptotic Mean ($A_n(t)$)</th>
<th>Bias (%)</th>
<th>Exact Standard Deviation ($A_n(t)$)</th>
<th>Asymptotic Standard Deviation ($A_n(t)$)</th>
<th>Bias (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size n=50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>100</td>
<td>49.936</td>
<td>49.934</td>
<td>0.002 (0.004%)</td>
<td>0.253</td>
<td>0.256</td>
<td>0.003 (1.001%)</td>
</tr>
<tr>
<td>0.001</td>
<td>500</td>
<td>49.606</td>
<td>49.598</td>
<td>0.008 (0.016%)</td>
<td>0.623</td>
<td>0.629</td>
<td>0.006 (0.981%)</td>
</tr>
<tr>
<td>0.001</td>
<td>1000</td>
<td>48.969</td>
<td>48.949</td>
<td>0.020 (0.042%)</td>
<td>0.994</td>
<td>1.004</td>
<td>0.009 (0.944%)</td>
</tr>
<tr>
<td>0.001</td>
<td>2000</td>
<td>46.371</td>
<td>46.302</td>
<td>0.069 (0.148%)</td>
<td>1.768</td>
<td>1.783</td>
<td>0.014 (0.797%)</td>
</tr>
<tr>
<td>0.003</td>
<td>100</td>
<td>49.929</td>
<td>49.927</td>
<td>0.001 (0.003%)</td>
<td>0.267</td>
<td>0.269</td>
<td>0.003 (1.001%)</td>
</tr>
<tr>
<td>0.003</td>
<td>500</td>
<td>49.299</td>
<td>49.285</td>
<td>0.014 (0.029%)</td>
<td>0.825</td>
<td>0.834</td>
<td>0.008 (0.963%)</td>
</tr>
<tr>
<td>0.003</td>
<td>1000</td>
<td>46.385</td>
<td>46.317</td>
<td>0.068 (0.148%)</td>
<td>1.765</td>
<td>1.780</td>
<td>0.014 (0.798%)</td>
</tr>
<tr>
<td>0.003</td>
<td>2000</td>
<td>18.999</td>
<td>18.679</td>
<td>0.319 (1.710%)</td>
<td>2.432</td>
<td>2.429</td>
<td>0.002 (0.097%)</td>
</tr>
<tr>
<td>0.010</td>
<td>100</td>
<td>49.895</td>
<td>49.893</td>
<td>0.002 (0.004%)</td>
<td>0.323</td>
<td>0.327</td>
<td>0.003 (0.999%)</td>
</tr>
<tr>
<td>0.010</td>
<td>300</td>
<td>48.858</td>
<td>48.835</td>
<td>0.023 (0.047%)</td>
<td>1.044</td>
<td>1.054</td>
<td>0.010 (0.937%)</td>
</tr>
<tr>
<td>0.010</td>
<td>600</td>
<td>33.502</td>
<td>33.266</td>
<td>0.236 (0.711%)</td>
<td>2.790</td>
<td>2.797</td>
<td>0.007 (0.248%)</td>
</tr>
<tr>
<td>0.010</td>
<td>800</td>
<td>10.919</td>
<td>10.582</td>
<td>0.337 (3.181%)</td>
<td>1.874</td>
<td>1.869</td>
<td>0.005 (0.259%)</td>
</tr>
</tbody>
</table>

Sample size n=200

Continued on next page
Next, to examine the performance of the asymptotic approximation over the entire time scale, we plot the mean and variance of the number of lineages as a function of time for two parameter settings in Fig. 3. In both settings, the contemporary population size and the sample size are chosen to be $N_0 = 2 \times 10^6$ and $n = 500$, but the growth rates are different: $\gamma = 0.003$ (Fig. 3 (A-B)) and $\gamma = 0.01$ (Fig. 3 (C-D)). We compare in Fig. 3 (A) and (C) three approaches that can be used to obtain the mean of the number of lineages: the sample mean from the simulated data, representing a close estimate of the true value, the exact mean as shown in Eqs. (24), and our proposed asymptotic mean. As shown in Fig. 3 (A) and (C), the asymptotic results well approximate the true mean of lineage numbers.

Recently, another approach to obtaining the expectation of the number of lineages was developed by
Figure 3: The mean and variance of the number of ancestral lineages in the history for two parameter settings.

(A) the mean of ancestral lineages as a function of time $t$ for $n = 500$ lineages sampled from the contemporary population. The contemporary population size is assumed to be $2 \times 10^6$, and the growth rate 0.003. The X-axis corresponds to generations back in time, and the Y-axis is the expectation of number of ancestral lineages. Green hollow circles represent the average of lineage numbers over 500 gene genealogies generated from coalescent simulations; the blue cross symbols represent the exact mean in Tavaré (1984); the red solid line represents the asymptotic mean derived in the main text.

(B) the variance of ancestral lineages as a function of time $t$ for $n = 500$ lineages sampled from the contemporary population. The parameters used in simulation and symbols are the same as (A) and the sample variance is estimated from 500 simulations.

(C) the mean of ancestral lineages as a function of time $t$ for $n = 500$ lineages sampled from the contemporary population. The contemporary population size is assumed to be $2 \times 10^6$, and the growth rate 0.01. The simulation setting and the representative symbols are the same as (A).

(D) the variance of ancestral lineages as a function of time $t$ for $n = 500$ lineages sampled from the contemporary population. The simulation setting and the representative symbols are the same as (A). The sample variance is estimated from 500 simulations.
MARUVKA et al. (2011) for populations with constant or exponentially growing size. Based on the equation for $E[A_n(t+1)|A_n(t) = i]$ in WATTERSON (1975), MARUVKA et al. (2011) constructed a differential equation and gave the solution for the NLFT (number of lineages as a function of time), referred as the expectation of the ancestral lineage number in this paper, as follows

$$E(A_n(t)) = \frac{n_0}{n_0 - (n_0 - 1)e^{-e^{\gamma t - 1}/2N_0\gamma}},$$  \hspace{1cm} (26)$$

where $n_0$ and $N_0$ are the current sample size and population size, and $\gamma$ is the population growth rate. Since MARUVKA et al. (2011) assumed $A_n(t)$ was deterministic, instead of a random variable, no formula for the variance of $A_n(t)$ was given in their paper. It can easily be shown that Eq. (26) is close to Eq. (21) when $A_n(t)$ is large. Letting $\beta = -\frac{1}{2}g(t) \rightarrow 0$, where $g(t) = \frac{e^{\gamma t - 1}}{N_0\gamma}$, the asymptotic mean of $A_n(t)$ in Eq. (21) tends to

$$\frac{n_0}{1 + \frac{n_0 - 1}{2N_0\gamma}}.$$  \hspace{1cm} (27)$$

When $g(t)$ is small, the denominator in Eq. (26) can be approximated by $n_0 - (n_0 - 1)(1 - \frac{e^{\gamma t - 1}}{2N_0\gamma}) = 1 + (n_0 - 1)\frac{e^{\gamma t - 1}}{2N_0\gamma}$, which is approximately the denominator in Eq. (27) when $n_0$ is large. This confirms the validity of our asymptotic approximation for exponential growth populations.

In Fig. 3 (B), (D), it is clearly evident that the asymptotic variance of $A_n(t)$ is close to the sample variance of the simulated data and the exact variance at any time $t$. We also notice that the variance of ancestral lineage numbers is on a relatively small magnitude compared to the expectation. This was exploited using simulation by MARUVKA et al. (2011) when they assumed that the number of ancestral lineages was nearly deterministic in large sample genealogies. However, the randomness of ancestral lineage numbers is still quite significant even for large sample genealogies. For example, for $n = 800, \gamma = 0.003$, the variance of $A_n(t)$ is around 100 at $t = 500$. If not taking into account of the randomness of ancestral lineages, the inference based on the coalescent likelihood will likely be biased. Our asymptotic results provide both the distribution and the analytical expressions of the two moments instead of only mean, and thus can be used to build statistically rigorous methods for parameter inference.

Finally, in addition to the mean and the variance, we also check how well the normal distribution approximates the exact distribution in shape using coalescent simulations. We examine the distribution of ancestral
Figure 4: The asymptotic probability density functions of lineage numbers in the history for two parameter settings.

(A), (C) and (E): Comparison of the asymptotic probability density function and simulated distribution of ancestral lineage number $A_n(t)$ for a population with $N_0 = 2 \times 10^6$ and growth rate being 0.003. The times $t$ are at 100 generations ago in (A), 1000 generations ago in (C) and 2000 generations ago in (E). The histograms were generated by 500 coalescent simulations.

(B), (D) and (F): Comparison of the asymptotic probability density function and simulated distribution of ancestral lineage number $A_n(t)$ for a population with $N_0 = 2 \times 10^6$ and growth rate being 0.01. The times $t$ are at 100 generations ago in (B), 400 generations ago in (D) and 800 generations ago in (F). The histograms were generated by 500 coalescent simulations.
lineage numbers at several time points for the same two parameter settings as in Fig. 3. We show snapshots at three time points for each setting in Fig. 4 as an illustration: for $\gamma = 0.003$, $t = 100, 1000$ and 2000 generations ago representing the early, middle and late stages of the ancestral process; and for $\gamma = 0.01$, $t = 100, 400$ and 800 generations ago. As can be seen from Fig. 4, the normal distribution provides a reasonable approximation to the true distribution of $A_n(t)$ for a wide range of time points.

We also examine how well the Poisson distribution and the gamma distribution approximate the distributions of coalesced lineages and coalescence times at the early stage of the coalescent process. The two asymptotic distributions only provide accurate approximations for the true distribution when the sample size $n$ is sufficiently large, $t$ is close to 0, and growth rate $\gamma$ is slow (see Tables S3 and S4, and Figs. S1 and S2 for details).

## 4 Applications

When the sample size $n$ is large, there exist numerical issues in evaluating the exact distributions of coalescence times and ancestral lineage numbers (Tavare, 1984; Griffiths and Tavare, 1998), and the asymptotic distributions derived above were shown to be a good approximation for finite sample sizes. Here we illustrate that the asymptotic distributions of coalescence times and the number of ancestral lineages can be applied to derive some fundamental statistics that summarize the properties of gene genealogies. We also show that the allele frequency spectrum of large-size samples can be derived through the asymptotic distribution of coalescence times for a population under exponential growth. These asymptotic statistics provide valid approximations and are in simple form without numerical issues for large samples.

### 4.1 Properties of large gene genealogies

Many statistics that summarize the properties of gene genealogies are informative for population genetic inference. Some of them can be derived as a function of coalescence times. We show the derivation of two important statistics of gene genealogies, the expected time to the most recent common ancestor (ETM-RCA) and the expected total branch length of the genealogy (ETBL). Using ETBL, we can easily estimate other summary statistics, such as Watterson’s diversity measure $\theta_W$ (Watterson, 1975) and Tajima’s D (Tajima, 1989).
Figure 5: The comparison of the expected TMRCA and total branch length obtained from the asymptotic distribution and the coalescent-simulated or exact distribution under three different exponential growth rates, assuming the contemporary population size to be $N_0 = 2 \times 10^6$.

(A) The expectation of time to the most recent common ancestors (ETMRCA) for different sample sizes. The curves correspond to the asymptotic ETMRCA for three growth rates ($\gamma = 0.001, 0.005, 0.01$). The hollowed circles, squares and diamonds are the averages of TMRCA over 200 coalescent simulations with respective growth rate.

(B) The expectation of total branch lengths of gene genealogies (ETBL) for different sample sizes. The curves correspond to the asymptotic ETBL for three growth rates ($\gamma = 0.001, 0.005, 0.01$). The hollowed circles, squares and diamonds are the exact ETBLs estimated using Eq. (35) of Polanski et al. (2003) with respective growth rates.
The time to the most recent common ancestor (TMRCA) of a sample is defined as the time when the ancestors of all lineages coalesce into a single ancient lineage, or the coalescence time $T_1$. Inferring TMRCA from genetic polymorphism data is of great interest in population genetic studies (Tavaré et al., 1997). ETMRCA by definition is simply the expectation of the coalescence time $T_1$:

$$\text{ETMRCA} = \mathbb{E}T_1. \quad (28)$$

The ETBL can be obtained by summing over the expectations of all branches of the genealogy as follows,

$$\text{ETBL} = \sum_{m=2}^{n} m\mathbb{E}W_m = \sum_{m=2}^{n} m \int_{0}^{\infty} \mathbb{P}(A_n(t) = m)dt$$
$$= \int_{0}^{\infty} \sum_{m=2}^{n} m\mathbb{P}(A_n(t) = m)dt$$
$$= \int_{0}^{\infty} \mathbb{E}(A_n(t))dt. \quad (29)$$

Specifically for populations under exponential growth, the ETMRCA can be approximated by

$$\text{ETMRCA} \approx \frac{1}{\gamma} \ln(2N_0\gamma(1 - n^{-1}) + 1), \quad (30)$$

and by substituting Eq. (27) into Eq. (29), we have

$$\text{ETBL} \approx \frac{2nN_0 \ln(2N_0\gamma)}{2N_0\gamma - n}. \quad (31)$$

In Fig. 5 (A), we show the ETMRCA as a function of sample size $n$ for three different growth rates ($\gamma = 0.001, 0.005$ and $0.01$) in a population with $N_0 = 2 \times 10^6$. The curves are theoretical predictions based on Eq. (30), and each point is an averaged TMRCA over 200 coalescent simulations for the sample size at $X$-axis. As we can see from the figure, the asymptotic ETMRCA is close to the simulated results, although biased towards larger values for slow growth rates (see Section 2.2 for the quantified bias). Given the large variance of TMRCA for genealogies under neutrality, the approximation is considerably accurate. When the growth rate increases, the theoretical approximation becomes more accurate. We can also see that the ETMRCA curve is nearly flat, which means that there is a limit as to how much the ETMRCA can
increase with the sample size. This is consistent with former conclusions that when the sample size is beyond a moderate level, adding more samples mainly changes the shape of lower parts of the gene genealogy and the increase in the height of the entire genealogy (TMRCA) is very minor (HEIN et al., 2005).

Estimating the ETBL directly from the exact intercoalescence time distributions suffers from severe numerical instability when the sample size is large. But as pointed out by POLANSKI et al. (2003), the estimation can be simplified by interchanging summations and cancelling large coefficient terms with each other in the alternating series, so that the resulting equations are computationally feasible for large samples (see Eq. (35) in POLANSKI et al. (2003)). We thus compare the asymptotic formula to the exact ETBL estimated from their equation. In application, we found that even with POLANSKI et al. (2003)’s technique, when the sample size is sufficiently large, the numerical instability issue still exists. For example, when \( n > 1600 \) for \( \gamma = 0.001 \), a high precision arithmetic library is needed to estimate the exact ETBL (Eq. (35) of POLANSKI et al. (2003)). The expected TBLs for the three levels of growth rates are shown in Fig. 5 (B). The asymptotic results fit the exact results very well. Furthermore, Eq. (31) is in simple and analytical form, making the computation for large samples much easier and faster than the exact equation in POLANSKI et al. (2003).

4.2 The allele frequency spectrum

The allele frequency spectrum (AFS) is defined as the sampling distribution of the frequency of mutant alleles in a randomly collected finite sample (CHEN, 2012). The AFS is informative for the inference of demographic history and natural selection (SAWYER and HARTL, 1992; WILLIAMSON et al., 2005; EVANS et al., 2007; CHEN et al., 2007; GUTENKUNST et al., 2009; LUKIĆ et al., 2011; ŽIVKOVIĆ and STEPHAN, 2011; SONG and STEINRÜCKEN, 2012; CHEN, 2012). It has been well studied using diffusion process since the foundation of population genetics (KIMURA, 1955). The AFS for any non-equilibrium populations can also be obtained in the coalescent framework using the expected time lengths of gene genealogies (GRIFFITHS and TAVARE, 1998; POLANSKI et al., 2003; MARTH et al., 2004; CHEN, 2012, 2013). The coalescent-based AFS was used extensively in studies of population growth, bottlenecks and other demographic history (WOODING and ROGERS, 2002; POLANSKI and KIMMEL, 2003; MARTH et al., 2004). The AFS under the coalescent model is derived in analytical form and is computationally efficient for small sam-
Figure 6: The theoretically predicted and simulated allele frequency spectrum for a sample of 500 lineages collected from exponentially-growing populations with different growth rates. The contemporary population size is $2 \times 10^6$ and the growth rates are (A) 0.001, (B) 0.005 and (C) 0.01 respectively. The three color bars in the histograms correspond to the simulated AFS, and the AFS estimated using the exact formulas of Polanski and Kimmel (2003) and the asymptotic formula derived in this paper respectively.
ples. However, the exact distribution of lineages is needed in deriving the AFS, which involves the sums of alternating series, and is difficult to evaluate for large samples. As a result, for large samples, either a high-precision arithmetic library has to be adopted (Wooding and Rogers, 2002; Marth et al., 2004), or the problem is transformed into a hypergeometric summation (Polanski and Kimmel, 2003). A high-precision arithmetic library requires tedious programming, and can significantly increase the computational time. The hypergeometric summation is a technique that allows efficient estimation of the AFS for large samples for any single population with temporally varying size. But it is not a general solution for the JAFS of two or multiple populations, and it is very challenging to extend the solution to other more complicated scenarios, such as migration and selection. Here we use the asymptotic distributions of coalescence times derived in Section 2 to get the approximation of the AFS for large samples. The derived AFS is in simple and analytical form and the approximation is accurate.

With the expectation of coalescence times derived in Eq. (8), we can easily get the expectation of intercoalescence times $\mathbb{E}W_m = \mathbb{E}T_{m-1} - \mathbb{E}T_m$. Let $\{\mathbb{E}S_j(n), 0 < j < n\}$ denote the AFS of SNPs from a sample of size $n$, with the $j$th entry $\mathbb{E}S_j(n)$ being the expected number of segregating sites having $j$ copies of derived alleles. The AFS can then be analytically estimated as in former studies (Fu, 1995; Griffiths and Tavaré, 1998):

$$
\mathbb{E}S_j(n) = \sum_{k=2}^{n} \frac{(n-j-1)!}{(n-1)!} \mu \times k \mathbb{E}(W_k) \times \frac{(n-j-1)!}{(n-1)!} \times \mu \sum_{k=2}^{n} k(k-1) \frac{n-k}{j-1} \mathbb{E}(W_k), 0 < j < n,
$$

where $\mu$ is the mutation rate per generation.

In Eq. (32), we assume an infinite-many-sites model for the mutation, so that mutations occurring at any of the $k$ branches spanning the time interval $(T_k, T_{k-1})$ follow a Poisson process with the mean of $\mu k \mathbb{E}(W_k)$. During the subsequent bifurcation process in which the number of lineages increases from $k$ to $n$, the count of the mutant increases from a single copy to $j$ among the $n$ lineages at present with the probability of $\frac{(n-j-1)!}{(n-1)!} \times \mu \sum_{k=2}^{n} k(k-1) \frac{n-k}{j-1} \mathbb{E}(W_k)$, which comes from the exchangability property among lineages.

The asymptotic AFS for large samples under variable population size model is estimated through the expected intercoalescence times by the above approach, with $\mathbb{E}(T_m)$ derived in Section 2.2. We present
three estimates of AFS in Fig. 6 to demonstrate the accuracy of asymptotic approximation: the sample AFS based on coalescent simulations, Polanski and Kimmel (2003)’s exact formula and the asymptotic AFS. The simulated samples of size 500 are assumed to come from a contemporary population with $N_0 = 2 \times 10^6$ and the exponential growth rate being 0.001, 0.005 and 0.01 respectively. The dark blue bars of the histograms in Fig. 6 are the AFS’s of a simulated 10Mb region, connected by 1000 regions of 10kb with the mutation rate and recombination rate both being $1 \times 10^{-8}$ per nucleotide. The light blue bars represent the exact AFS estimated using Polanski and Kimmel (2003)’s exact formulas (Eqs. (8-15) of their paper), and the white bars represent the asymptotic AFS calculated based on Eq. (32) for the above chosen parameters. For a zoomed-in view, we only present the first 25 entries of the AFS’s in Fig. 6. The AFS based on the asymptotic distribution of intercoalescence times matches both the exact result and the simulation result accurately. The asymptotic result derived in this paper has advantages over the existing methods. It is in simple analytical form, and the calculation is fast without numerical instability and doesn’t involve numerical integral or sampling-based methods. Also, the method can be flexibly generalized to various population history models other than the simple exponential growth model. The extension of the asymptotic AFS to various demographic history models, especially the joint AFS for multiple populations, will be be investigated in the future studies.

5 Discussion

The distributions of coalescence times and the number of ancestral lineages play an essential role in coalescent modeling and population genetic inference. Both exact distributions of ancestral lineage numbers and coalescence times have been studied and expressed as a sum of alternating series, the terms of which are difficult to evaluate when the sample size is large (Tavaré, 1984; Griffiths and Tavaré, 1998; Polanski et al., 2003). With the rapid advancement of sequencing technology, large-sample genomic sequencing data are piling up, calling for new coalescent theories and methods for population genetic analysis. This paper extends the asymptotic distributions of ancestral lineage numbers and coalescence times in constant populations (Griffiths, 1984) to populations with temporally varying size. The asymptotic distributions provide a computationally fast and reliable alternative to the exact distributions in large samples. And we have shown that the asymptotic distributions are useful to obtain statistics describing the properties of
large genealogies, and to analytically construct the large-sample allele frequency spectrum. We expect the theoretical results derived in this paper, together with the results in Griffiths (1984), to be useful for coalescent-based methodology development at the age of population-level sequencing data.

Acknowledgement

We are grateful to Dr. Robert Griffiths for insightful comments on an earlier version of the manuscript, which greatly improved the work. We are also grateful to Dr. Joachim Hermisson and the two anonymous reviewers for their helpful comments.

References


