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

Hua Chen^{*.1} and Kun Chen[†]

^{*}Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China, [†]Dana Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115

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 article, 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 $t \rightarrow 0$, the number of coalesced lineages $n - A_n(t)$ follows a Poisson distribution, and as $m \rightarrow n$, $n(n - 1)T_m/2N(0)$ 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.

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 article, 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. 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 (Risch *et al.* 2003; Anderson and Slatkin 2007; 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 (Equation 15 under *Asymptotics of ancestral lineage numbers*; see also Griffiths

Copyright © 2013 by the Genetics Society of America

doi: 10.1534/genetics.113.151522

Manuscript received April 4, 2013; accepted for publication May 4, 2013

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.151522/-/DC1>.

¹Corresponding author: Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China. E-mail: hchen007@gmail.com

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 $k(k - 1)/2$, $k = n, \dots, m + 1$ under a constant population size model. The analytical expressions of many statistics are derived on the basis of 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; Altshuler *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 applicable only 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 on the basis of 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 numeri-

cal 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 $t \rightarrow 0$ and the sample size $n \rightarrow \infty$, 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 $\int_0^t (ds/N(s))$, $t \geq 0$, 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 nonequilibrium 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 *Asymptotic Distributions for Coalescence Times and Ancestral Lineage Numbers* 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 *Numerical Results* 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 *Applications*, 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 AFS for large samples in simpler analytical form. The article closes with a discussion.

Asymptotic Distributions for Coalescence Times and Ancestral Lineage Numbers

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 Kingman's coalescent assumption ($N(t) \gg n$). 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 investigate 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 $T_n = 0$ (Figure 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 $m(m-1)/2N_0$ 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 nonmutant 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 $\tau = g(t) = \int_0^t (1/N(u)) du$ be the time scaled at rate $N(t)$, specifically for populations with constant size, $\tau = t/N_0$. We use the notations $\dot{T}_m$, $\dot{W}_m$, and $\dot{A}_n(\tau)$ 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 $n \rightarrow \infty$, $n/m \rightarrow a$, $1 < a < \infty$, the m th coalescence time, T_m , is asymptotically normal,

Figure 1 An illustration of the gene genealogy and coalescence times of five lineages at present. The coalescence time T_m is defined as the time when $m + 1$ lineages coalesce into m lineages.

with the mean and variance depending on the historical population size (Equations 8 and 9);

2. as $g(t) \rightarrow 0$, $n \rightarrow \infty$, $\frac{1}{2}ng(t) \rightarrow \alpha$, $0 < \alpha \leq \infty$, the number of ancestral lineages at time t , $A_n(t)$, is asymptotically normal, with the mean and variance provided in Equations 19 and 20; and
3. at the very early stage of the coalescent, when $t \rightarrow 0$ more rapidly, $n \rightarrow \infty$, the number of coalesced lineages $n - A_n(t)$ follows a Poisson distribution, and when $m \rightarrow n$, $n - m$ is bounded above, $(n(n-1)T_m)/2N_0$ follows a gamma distribution.

Asymptotics of coalescence times

Under a constant population-size model, the intercoalescence times, W_m , $2 \leq m \leq n$, are independent exponential variables with respective rates $\binom{m}{2}$. 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 $\dot{T}_m$ 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 $n \rightarrow \infty$, $m \rightarrow \infty$, while $n/m \rightarrow a$, $1 < a \leq \infty$, $\dot{T}_m$ is asymptotically normal, and the mean and the variance of $\dot{T}_m$ are

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

and

$$\begin{aligned} \sigma_m^2 &= \sum_{k=m+1}^n \frac{4}{k^2(k-1)^2} = 4 \sum_{k=m+1}^n \left[\frac{1}{(k-1)^2} + \frac{1}{k^2} - \frac{2}{k(k-1)} \right] \\ &= 4 \{ \psi_1(m) - \psi_1(n) + \psi_1(m+1) - \psi_1(n+1) - 2(m^{-1} - n^{-1}) \}, \end{aligned} \quad (3)$$

where the trigamma function $\psi_1(z) = (d^2/dz^2) \ln \Gamma(z)$ 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 Equations 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 $(T_m, \dots, T_{n-1})$ for populations with temporally varying size is (Griffiths and Tavaré 1998):

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

The marginal probability density function (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 on the basis of the definition of the n -coalescent and a pure-death process (see Griffiths 2006 and Chen 2012, Appendix A, for details of derivation):

$$f_{T_m}(t) = \mathbb{P}(A_n(t) = m+1) \frac{(m+1)m}{2N(t)}. \quad (5)$$

As $\mathbb{P}(A_n(t) = m+1)$ was derived from the expansion of the transition function (see Equations 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 $1/N(t)$, $g(T_m) = \int_0^{T_m} (1/N(u)) du$, follows the distribution of coalescence time in the standard Kingman's n -coalescent, and the scaled intercoalescence times,

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

are mutually independent exponential variables with the rate of $\binom{m}{2}$. The coalescent process under a variable population size model, as sample size $n \rightarrow \infty$, is still Kingman's coalescent since we assume the population size tends to infinity more quickly than the sample size, in other words, $n/N(t) \rightarrow 0$, 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 $g^{-1}(\mu_m)$,

$$g(T_m) = \mu_m + g'(g^{-1}(\mu_m)) (T_m - g^{-1}(\mu_m)) + \frac{g''(g^{-1}(\mu_m))}{2} (T_m - g^{-1}(\mu_m))^2 + O((T_m - g^{-1}(\mu_m))^3), \quad (6)$$

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 Equation 6 is ignorable as $n/m \rightarrow a$, $n \rightarrow \infty$, because $g(T_m)$ follows the same distribution as $\hat{T}_m$ and $\hat{T}_m \rightarrow \mu_m$ by the asymptotic properties of $\hat{T}_m$ shown in Griffiths (1984, Theorem 1).

Next, by Equation 6 and ignoring the remainder term, we have

$$\frac{T_m - g^{-1}(\mu_m)}{(\sigma_m/g'(g^{-1}(\mu_m)))} \rightarrow \frac{g(T_m) - \mu_m}{\sigma_m}. \quad (7)$$

As $(g(T_m) - \mu_m)/\sigma_m \rightarrow N(0, 1)$, the limiting distribution of T_m can then be approximated by a normal distribution with the mean

$$\mathbb{E}(T_m) = g^{-1}(\mu_m), \quad (8)$$

and variance

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

Substituting Equations 2 and 3 into Equations 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 γ , that is, $N(t) = N_0 e^{-\gamma t}$, it is straightforward to write the scaling function as

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

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

$$\begin{aligned} \mathbb{E}(T_m) &= g^{-1}(\mu_m) = \frac{\ln(N_0 \gamma \mu_m + 1)}{\gamma} \\ &= \frac{1}{\gamma} \ln(2N_0 \gamma (m^{-1} - n^{-1}) + 1), \end{aligned} \quad (11)$$

and variance

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

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,

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

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

$$\begin{aligned}
\mathbb{E}(T_m) - g^{-1}(\mu_m) &= \frac{(g^{-1})''(\mu_m)}{2} \mathbb{E}(g(T_m) - \mu_m)^2 + o(1) \\
&= -\frac{N_0^2 \gamma \sigma_m^2}{2(N_0 \gamma \mu_m + 1)^2} + o(1)
\end{aligned} \tag{13}$$

$$= -\frac{\sigma_m^2}{2\gamma\mu_m^2} + o(1). \tag{14}$$

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

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; Watterson 1984; Takahata and Nei 1985),

$$\mathbb{P}(A_n(t) = m) = \sum_{i=m}^n \frac{(-1)^{i-m} (2i-1) m_{(i-1)} n_{[i]}}{m!(i-m)!n_{(i)}} e^{-i(i-1)t/2N}, \quad 0 < m \leq n, \tag{15}$$

where $n_{(i)} = n(n+1)\dots(n+i-1), i \geq 1; n_{(0)} = 1$, and $n_{[i]} = n(n-1)\dots(n-i+1), i \geq 1; n_{[0]} = 1$ are the rising and falling factorial functions. Griffiths and Tavaré (1998)

further generalized Equation 15 to populations with variable size. Then for populations with temporally varying size, the distribution of the number of ancestral lineages becomes (Griffiths and Tavaré 1998):

$$\mathbb{P}(A_n(t) = m) = \sum_{i=m}^n \frac{(-1)^{i-m} (2i-1) m_{(i-1)} n_{[i]} e^{[-i(i-1)]/2N_0 \int_0^t (1/\lambda(u)) du}}{m!(i-m)!n_{(i)}}. \tag{16}$$

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

$$\mu(\tau) = \frac{2\eta}{\tau}, \tag{17}$$

and variance

$$\sigma^2(\tau) = 2\eta\tau^{-1}(\eta + \beta)^2 \{1 + \eta/(\eta + \beta) - \eta/\alpha - \eta/(\alpha + \beta) - 2\eta\}\beta^{-2}, \tag{18}$$

where $\beta = -\frac{1}{2}\tau, \eta = \alpha\beta/\{\alpha(e^\beta - 1) + \beta e^\beta\}$. In the following, we extend Griffiths' 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 Griffiths' Theorem 2, as $n \rightarrow \infty, \tau \rightarrow 0$ and $\frac{1}{2}n\tau \rightarrow \alpha$, the ancient lineage number of the coalescent process $\dot{A}_n(\tau) \rightarrow_d N(u(\tau), \sigma^2(\tau))$, where $\mu(\tau)$ and $\sigma^2(\tau)$ are given in Equations 17 and 18. Therefore, $\mathbb{P}(A_n(t) \leq m) \rightarrow \mathbb{P}(Z \leq (m - u(\tau))/\sigma(\tau))$, 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

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

and

$$\begin{aligned}
\text{Var}(A_n(t)) &\rightarrow \sigma^2(\tau) = \sigma^2(g(t)) \\
&= 2\eta(g(t))^{-1}(\eta + \beta)^2 \{1 + \eta/(\eta + \beta) - \eta/\alpha - \eta/(\alpha + \beta) - 2\eta\}\beta^{-2},
\end{aligned} \tag{20}$$

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

When the population is under exponential growth with rate γ and scaling function as in Equation 10, plugging Equation 10 into Equations 19 and 20, $A_n(t)$ then has an asymptotic mean

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

and variance

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

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

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 $A_n(t)$.

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

At the early stage of the coalescent, or $m \rightarrow n$ and $t \rightarrow 0$, 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 $n(n-1)/2$, $[n(n-1)/2](g(T_m) - g(T_{m+1})) \sim \text{Exponential}(\binom{m+1}{2}/\binom{n}{2}) \rightarrow \text{Exponential}(1)$ as $m \rightarrow n$ 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 $g'(\xi_{m+1}) = 1/N(\xi_{m+1}) \rightarrow 1/N_0$. Subsequently,

$$\frac{n(n-1)}{2} \sum_{k=m+1}^n g(T_{k-1}) - g(T_k) = \frac{n(n-1)}{2} \sum_{k=m+1}^n (T_{k-1} - T_k)g'(\xi_k). \quad (23)$$

Taking limits $n \rightarrow \infty$, $m \rightarrow n$ at both sides of Equation 23, we obtain $[n(n-1)/2N_0]T_m \rightarrow \gamma(n-m, 1)$.

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 $\mathbb{P}(n - A_n(t) \geq n - j) = \mathbb{P}(n - \dot{A}_n(g(t)) \geq n - j) = \mathbb{P}(n - \dot{A}_n(\tau) \geq n - j)$, and Theorem 6 in Griffiths (1984), $n - A_n(t)$ asymptotically follows the Poisson distribution with mean $\nu = \frac{1}{2}n(n-1)g(t) = \frac{1}{2}n(n-1) \int_0^t [1/N(u)]du$.

Numerical Results

The accuracy of the asymptotic distributions of T_m and $A_n(t)$ derived in *Asymptotic Distributions for Coalescence Times and Ancestral Lineage Numbers* 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 2×10^6 , and a wide range of exponential growth rate γ , time t , and sample size n is investigated.

Coalescence times

We first examine the numerical accuracy of the asymptotic p.d.f. of coalescence times as provided in *Asymptotics of coalescence times*. We show the asymptotic p.d.f.'s of several coalescence times for $n = 500$ and $\gamma = 0.001$ and 0.005 in Figure 2, and then present in Table 1 the mean and standard deviation of coalescence times for a wider range of parameters (more simulation results can be seen in [Supporting Information, Table S1](#)). Since evaluating the p.d.f. of coalescence times for large samples using the exact formulas is subject to severe numerical instability (Polanski *et al.* 2003), we use only coalescent simulations in this section.

In Figure 2, A and B, the asymptotic p.d.f.'s of coalescence times, $T_2, T_{10}, T_{25}, T_{50}, T_{100}, T_{200}, T_{300}$, and T_{400} , are displayed for two different exponential growth rates, 0.001 and 0.005 . It is obvious that coalescence times on average are longer in a slowly growing population, so that the gene genealogy of a faster-growing population tends to have relatively shorter internal branches, showing a "star" shape. In Figure 2, C and E, the p.d.f.'s of T_{10} and T_{100} under growth model $\gamma = 0.001$ are presented together with the respective histograms generated from 500 coalescent simulations. A similar comparison was shown for $\gamma = 0.005$ in Figure 2, D and F. The asymptotic p.d.f. is reasonably close to the simulated distribution for the two selected T_m 's and the two growth rates.

In Table 1 and [Table 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 Table 1 and [Table S1](#) the asymptotic mean of T_m is always larger than the simulated mean. This is consistent with our quantified bias in Equation 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 μ_m close to 0 (approximately $-1/\gamma\mu_m^2$; see also Equation 14) appearing in the bias term. The detailed derivation of the quantified bias can be seen in *Asymptotics of coalescence times*.

Another trend worth noting in Table 1 and [Table S1](#) are that the relative bias of the mean decreases with

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 (A) $\gamma = 0.001$ and (B) $\gamma = 0.005$. 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 (C) 0.001 and (D) 0.005. 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 (E) 0.001 and (F) 0.005. The histograms were generated by 500 coalescent simulations.

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 Table 1 and Table S1). For a coalescence time T_m with a smaller m , the relative bias is reduced more slowly. In Table 1 and Table S1, the relative bias for T_5 does not have obvious trend of decrease for sample sizes

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

γ	m	Mean (T_m)			Standard deviation (T_m)		
		Simulation	Asymptotic	Bias (%)	Simulation	Asymptotic	Bias (%)
Sample size $n = 50$							
0.001	5	6538.619	6580.639	42.020 (0.643%)	278.445	285.229	6.784 (2.436%)
0.001	10	5746.244	5771.441	25.197 (0.438%)	223.672	226.368	2.696 (1.206%)
0.001	20	4775.353	4795.791	20.438 (0.428%)	209.097	206.392	2.705 (1.294%)
0.001	45	2210.726	2291.412	80.685 (3.650%)	402.151	402.703	0.552 (0.137%)
0.005	5	1629.978	1637.793	7.816 (0.480%)	56.458	57.109	0.651 (1.154%)
0.005	10	1470.136	1475.677	5.541 (0.377%)	44.957	45.387	0.429 (0.955%)
0.005	20	1274.626	1279.719	5.093 (0.400%)	41.512	41.553	0.041 (0.100%)
0.005	45	742.147	763.298	21.151 (2.850%)	90.637	87.630	3.008 (3.318%)
0.010	5	884.172	888.198	4.026 (0.455%)	28.072	28.559	0.486 (1.731%)
0.010	10	804.606	807.122	2.515 (0.313%)	22.611	22.700	0.089 (0.395%)
0.010	20	706.889	709.091	2.202 (0.312%)	20.902	20.794	0.108 (0.517%)
0.010	45	439.281	449.857	10.577 (2.408%)	46.789	44.302	2.486 (5.314%)
Sample size $n = 200$							
0.001	5	6626.114	6660.575	34.462 (0.520%)	254.189	263.447	9.258 (3.642%)
0.001	20	5187.985	5198.497	10.512 (0.203%)	141.164	142.544	1.380 (0.978%)
0.001	50	4104.733	4110.874	6.141 (0.150%)	105.829	106.237	0.408 (0.386%)
0.001	100	3038.662	3044.522	5.861 (0.193%)	103.844	102.868	0.977 (0.940%)
0.001	150	2028.941	2036.882	7.941 (0.391%)	126.889	124.671	2.219 (1.749%)
0.001	195	405.591	413.976	8.385 (2.067%)	147.663	151.613	3.950 (2.675%)
0.005	5	1647.798	1653.798	6.000 (0.364%)	51.114	52.743	1.629 (3.187%)
0.005	20	1359.062	1360.701	1.639 (0.121%)	28.242	28.635	0.393 (1.392%)
0.005	50	1140.012	1141.422	1.410 (0.124%)	21.275	21.530	0.255 (1.197%)
0.005	100	921.898	923.024	1.126 (0.122%)	21.349	21.388	0.039 (0.184%)
0.005	150	704.509	707.223	2.715 (0.385%)	27.911	27.839	0.071 (0.256%)
0.005	195	243.791	254.182	10.391 (4.262%)	62.681	64.354	1.672 (2.668%)
0.010	5	893.062	896.201	3.139 (0.352%)	25.688	26.375	0.687 (2.676%)
0.010	20	748.544	749.610	1.066 (0.142%)	14.058	14.326	0.268 (1.905%)
0.010	50	639.188	639.859	0.671 (0.105%)	11.055	10.783	0.273 (2.466%)
0.010	100	529.684	530.330	0.646 (0.122%)	10.860	10.747	0.112 (1.033%)
0.010	150	420.709	421.459	0.751 (0.178%)	14.020	14.125	0.105 (0.749%)
0.010	195	174.898	181.290	6.392 (3.655%)	37.343	37.428	0.085 (0.227%)
Sample size $n = 800$							
0.001	5	6645.512	6679.599	34.087 (0.513%)	247.504	258.485	10.980 (4.436%)
0.001	10	5958.249	5981.414	23.165 (0.389%)	181.338	184.235	2.897 (1.598%)
0.001	50	4328.359	4330.733	2.374 (0.055%)	85.882	85.933	0.051 (0.059%)
0.001	200	2771.704	2772.589	0.885 (0.032%)	50.952	50.631	0.321 (0.630%)
0.001	400	1789.988	1791.759	1.771 (0.099%)	44.957	45.005	0.048 (0.106%)
0.001	795	30.645	30.962	0.317 (1.036%)	13.477	13.635	0.157 (1.167%)
0.005	5	1651.462	1657.606	6.144 (0.372%)	50.473	51.749	1.276 (2.528%)
0.005	10	1515.254	1517.766	2.512 (0.166%)	36.394	36.922	0.528 (1.451%)
0.005	50	1185.082	1185.918	0.835 (0.070%)	17.362	17.369	0.007 (0.042%)
0.005	200	865.872	866.147	0.275 (0.032%)	10.690	10.659	0.031 (0.291%)
0.005	400	651.296	651.619	0.323 (0.050%)	10.458	10.386	0.072 (0.689%)
0.005	795	28.726	29.206	0.481 (1.673%)	12.000	12.153	0.153 (1.274%)
0.010	5	894.511	898.105	3.594 (0.402%)	25.293	25.878	0.585 (2.314%)
0.010	10	825.938	828.172	2.234 (0.270%)	18.054	18.465	0.412 (2.281%)
0.010	50	661.756	662.141	0.385 (0.058%)	8.640	8.696	0.056 (0.648%)
0.010	200	501.546	501.728	0.182 (0.036%)	5.500	5.365	0.135 (2.449%)
0.010	400	393.125	393.183	0.058 (0.015%)	5.358	5.295	0.064 (1.188%)
0.010	795	26.667	27.343	0.676 (2.534%)	10.423	10.699	0.276 (2.652%)

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 Table 1 and Table S1). Although the convergence rates are different for the above two T_m 's, the bias of both T_m 's shrinks toward zero as $n \rightarrow \infty$.

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),

$$\mathbb{E}^{\text{exact}}(A_n(t)) = \sum_{l=1}^n (2l-1) \frac{n_{[l]}}{n_{(l)}} \exp\left\{-\frac{l(l-1)}{2N_0} \int_0^t \frac{1}{\lambda(u)} du\right\}, \quad (24)$$

and

$$\mathbb{E}^{\text{exact}}(A_n^2(t)) = \sum_{l=1}^n (2l-1) \frac{(l(l-1)+1) \times n_{[l]}}{n_{(l)}} \exp\left\{-\frac{l(l-1)}{2N_0} \int_0^t \frac{1}{\lambda(u)} du\right\}. \quad (25)$$

Unlike the entire exact distribution of the ancestral lineage number, the exact mean and variance of ancestral lineage numbers can be accurately calculated from Equations 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 $\gamma = 0.01$).

Table 2 and Table 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 Table 2 and Table 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.

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 Figure 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: $\gamma = 0.003$ (Figure 3, A–B) and $\gamma = 0.01$ (Figure 3, C–D). We compare in Figure 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 Equations 24, and our proposed asymptotic mean. As shown in Figure 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 Maruvka *et al.* (2011) for populations with constant or exponentially grow-

ing size. Based on the equation for $\mathbb{E}[A_n(t+1)|A_n(t) = i]$ in Watterson (1975), Maruvka *et al.* (2011) constructed a differential equation and gave the solution for the number of lineages as a function of time (NLFT), referred as the expectation of the ancestral lineage number in this article, as follows

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

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

$$\frac{n_0}{1 + [n_0 \cdot (e^{\gamma t} - 1)]/2N_0\gamma}. \quad (27)$$

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

In Figure 3, B and 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 note 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 ~ 100 at $t = 500$. If not taking into account 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 check how well the normal distribution approximates the exact distribution in shape using coalescent simulations. We examine the distribution of ancestral lineage numbers at several time points for the same two parameter settings as in Figure 3. We show snapshots at three time points for each setting in Figure 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 Figure 4, the normal distribution provides a reasonable approximation to the true distribution of $A_n(t)$ for a wide range of time points.

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

		Mean ($A_n(t)$)			Standard deviation ($A_n(t)$)		
γ	t	Exact	Asymptotic	Bias (%)	Exact	Asymptotic	Bias (%)
Sample size $n = 50$							
0.001	100	49.936	49.934	0.002 (0.004%)	0.253	0.256	0.003 (1.001%)
0.001	500	49.606	49.598	0.008 (0.016%)	0.623	0.629	0.006 (0.981%)
0.001	1000	48.969	48.949	0.020 (0.042%)	0.994	1.004	0.009 (0.944%)
0.001	2000	46.371	46.302	0.069 (0.148%)	1.768	1.783	0.014 (0.797%)
0.003	100	49.929	49.927	0.001 (0.003%)	0.267	0.269	0.003 (1.001%)
0.003	500	49.299	49.285	0.014 (0.029%)	0.825	0.834	0.008 (0.963%)
0.003	1000	46.385	46.317	0.068 (0.148%)	1.765	1.780	0.014 (0.798%)
0.003	2000	18.999	18.679	0.319 (1.710%)	2.432	2.429	0.002 (0.097%)
0.010	100	49.895	49.893	0.002 (0.004%)	0.323	0.327	0.003 (0.999%)
0.010	300	48.858	48.835	0.023 (0.047%)	1.044	1.054	0.010 (0.937%)
0.010	600	33.502	33.266	0.236 (0.711%)	2.790	2.797	0.007 (0.248%)
0.010	800	10.919	10.582	0.337 (3.181%)	1.874	1.869	0.005 (0.259%)
Sample size $n = 200$							
0.001	100	198.959	198.954	0.005 (0.003%)	1.015	1.017	0.003 (0.246%)
0.001	500	193.747	193.717	0.030 (0.016%)	2.423	2.428	0.006 (0.227%)
0.001	1000	184.250	184.177	0.073 (0.040%)	3.660	3.667	0.007 (0.195%)
0.001	2000	151.766	151.578	0.188 (0.124%)	5.336	5.341	0.005 (0.102%)
0.003	100	198.846	198.841	0.006 (0.003%)	1.068	1.071	0.003 (0.246%)
0.003	500	189.083	189.031	0.052 (0.027%)	3.125	3.132	0.007 (0.211%)
0.003	1000	151.922	151.734	0.188 (0.124%)	5.332	5.338	0.005 (0.102%)
0.003	2000	26.285	25.950	0.335 (1.292%)	2.941	2.938	0.004 (0.121%)
0.010	100	198.305	198.296	0.008 (0.004%)	1.291	1.294	0.003 (0.244%)
0.010	300	182.657	182.577	0.080 (0.044%)	3.808	3.816	0.007 (0.189%)
0.010	600	66.720	66.398	0.322 (0.485%)	4.619	4.618	0.002 (0.037%)
0.010	800	12.917	12.579	0.339 (2.692%)	2.052	2.047	0.005 (0.236%)
Sample size $n = 800$							
0.001	100	783.539	783.519	0.020 (0.003%)	3.974	3.976	0.002 (0.059%)
0.001	500	708.227	708.125	0.102 (0.014%)	8.502	8.505	0.004 (0.043%)
0.001	1000	595.586	595.390	0.196 (0.033%)	10.798	10.801	0.003 (0.024%)
0.001	2000	351.520	351.214	0.305 (0.087%)	10.352	10.352	0.000 (0.003%)
0.003	100	781.788	781.766	0.022 (0.003%)	4.171	4.173	0.002 (0.058%)
0.003	500	649.446	649.291	0.155 (0.024%)	10.033	10.036	0.003 (0.032%)
0.003	1000	352.361	352.055	0.305 (0.087%)	10.361	10.361	0.000 (0.003%)
0.003	2000	29.083	28.747	0.336 (1.168%)	3.099	3.095	0.003 (0.111%)
0.010	100	773.453	773.421	0.032 (0.004%)	4.982	4.985	0.003 (0.056%)
0.010	300	579.199	578.992	0.207 (0.036%)	10.944	10.947	0.002 (0.021%)
0.010	600	88.745	88.412	0.334 (0.377%)	5.427	5.425	0.002 (0.037%)
0.010	800	13.540	13.202	0.338 (2.564%)	2.102	2.098	0.005 (0.227%)

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 provide only accurate approximations for the true distribution when the sample size n is sufficiently large, t is close to 0, and growth rate γ is slow (see Table S3, Table S4, Figure S1, and Figure S2 for details).

Applications

When the sample size n is large, numerical issues in evaluating the exact distributions of coalescence times and ancestral lineage numbers exist (Tavaré 1984; Griffiths and Tavaré 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.

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

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 2×10^6 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 open circles represent the average of lineage numbers over 500 gene genealogies generated from coalescent simulations; the blue X 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 in 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 2×10^6 and the growth rate 0.01. The simulation setting and the representative symbols are the same as in 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 in A. The sample variance is estimated from 500 simulations.

coalescence times. We show the derivation of two important statistics of gene genealogies, the expected time to the most recent common ancestor (ETMRCA) 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 θ_W (Watterson 1975) and Tajima's D (Tajima 1989).

The time to 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 are 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

$$\begin{aligned} \text{ETBL} &= \sum_{m=2}^n m \mathbb{E}W_m = \sum_{m=2}^n m \int_{t=0}^{\infty} \mathbb{P}(A_n(t) = m) dt \\ &= \int_0^{\infty} \sum_{m=2}^n m \mathbb{P}(A_n(t) = m) dt \\ &\approx \int_0^{\infty} \mathbb{E}(A_n(t)) dt. \end{aligned} \quad (29)$$

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 Equation 27 into Equation 29, we have

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 0.003. The times t are at (A) 100 generations ago, (C) 1000 generations ago, and (E) 2000 generations ago. 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 (B) 100 generations ago, (D) 400 generations ago, and (F) 800 generations ago. The histograms were generated by 500 coalescent simulations.

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

In Figure 5A, 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 Equation 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 toward larger values for slow

growth rates (see *Asymptotics of coalescence times* 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 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 canceling large coefficient terms with each other in the alternating series, so that the resulting equations are computationally feasible for large samples (see Polanski *et al.* 2003, Equation 35). 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 (Polanski *et al.* 2003, Equation 35). The expected TBLs for the three levels of growth rates are shown in Figure 5B. The asymptotic results fit the exact results very well. Furthermore, Equation 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).

The allele frequency spectrum

The 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 nonequilibrium populations can also be obtained in the coalescent framework using the expected time lengths of gene genealogies (Griffiths and Tavaré 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 samples. However, the exact distribution of lineages is needed in deriving the AFS, which involves the sums of alternating series, and is difficult to

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 open 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 open circles, squares, and diamonds are the exact ETBLs estimated using Equation 35 of Polanski *et al.* (2003) with respective growth rates.

evaluate for large samples. As a result, for large samples, either a high-precision arithmetic library must 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 *Asymptotic*

Distributions for Coalescence Times and Ancestral Lineage Numbers 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 Equation 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):

$$\begin{aligned} \mathbb{E} S_j(n) &= \sum_{k=2}^n \left[\binom{n-j-1}{k-2} / \binom{n-1}{k-1} \right] \mu \times k \mathbb{E}(W_k) \\ &= [(n-j-1)!(j-1)!/(n-1)!] \mu \sum_{k=2}^n k(k-1) \\ &\quad \times \binom{n-k}{j-1} \mathbb{E}(W_k), 0 < j < n, \end{aligned} \quad (32)$$

where μ is the mutation rate per generation.

In Equation 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

$$\frac{\binom{n-j-1}{k-2}}{\binom{n-1}{k-1}},$$

which comes from the exchangeability 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 *Asymptotics of coalescence times*. We present three estimates of AFS in Figure 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 Figure 6 are the AFSs of a simulated 10-Mb region, connected by 1000 regions of 10 kb with the mutation rate and recombination rate both being 1×10^{-8} /nucleotide. The light blue bars represent the exact AFS estimated using Polanski and Kimmel (2003)'s exact formulas (Equations 8–15 in their article), and the white bars represent the asymptotic AFS calculated based on Equation 32 for the above chosen parameters. For a zoomed-in view, we present only the first 25 entries of the AFSs in Figure 6.

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 2×10^6 and the growth rates are (A) 0.001, (B) 0.005, and (C) 0.01. 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 article, respectively.

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

article 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 investigated in the future studies.

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 article 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 in obtaining statistics describing the properties of large genealogies and in analytically constructing the large-sample allele frequency spectrum. We expect the theoretical results derived in this article, together with the results in Griffiths (1984), to be useful for coalescent-based methodology development at the age of population-level sequencing data.

Acknowledgments

We are grateful to Dr. Robert Griffiths for insightful comments on an earlier version of the manuscript, which greatly improved the work. We are grateful to Dr. Joachim Hermisson and the two anonymous reviewers for their helpful comments. We are also grateful to Drs. Li Jin, Bing Su, and Hong Shi for motivating and encouraging the work.

Literature Cited

- Altshuler, D., E. Lander, L. Ambrogio, T. Bloom, K. Cibulskis *et al.*, 2010 A map of human genome variation from population scale sequencing. *Nature* 467: 1061–1073.
- Anderson, E., and M. Slatkin, 2007 Estimation of the number of individuals founding colonized populations. *Evolution* 61: 972–983.
- Billingsley, P., 2012 *Probability and Measure*. Wiley, New York.
- Chen, H., 2012 The joint allele frequency spectrum of multiple populations: a coalescent theory approach. *Theor. Popul. Biol.* 81: 179–195.
- Chen, H., 2013 Intercoalescence time distribution of incomplete genealogies in temporally varying populations, and applications in population genetic inference. *Ann. Hum. Genet.* 77: 158–173.
- Chen, H., R. E. Green, S. Pääbo, and M. Slatkin, 2007 The joint allele-frequency spectrum in closely related species. *Genetics* 177: 387–398.
- Coventry, A., L. M. Bull-Ottersson, X. Liu, A. G. Clark, T. J. Maxwell *et al.*, 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* 1: 131.
- Dlugosch, K., and I. Parker, 2007 Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Mol. Ecol.* 17: 431–449.
- Donnelly, P., 1984 The transient behaviour of the moran model in population genetics. *Math. Proc. Camb. Philos. Soc.* 95: 349–358.
- Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29: 401–421.
- Evans, S., Y. Shvets, and M. Slatkin, 2007 Non-equilibrium theory of the allele frequency spectrum. *Theor. Popul. Biol.* 71: 109–119.
- Ewens, W., 1972 The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3: 87–112.
- Ewens, W., 2004 *Mathematical Population Genetics: Theoretical Introduction*, Vol. 1. Springer Verlag, New York.
- Felsenstein, J., M. Kuhner, J. Yamato, and P. Beerli, 1999 Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. *Lect. Notes Monogr. Ser.*, 163–185.
- Fu, Y. X., 1995 Statistical properties of segregating sites. *Theor. Popul. Biol.* 48: 172–197.
- Griffiths, R. C., 1980 Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theor. Popul. Biol.* 17: 37–50.
- Griffiths, R. C., 1984 Asymptotic line-of-descent distributions. *J. Math. Biol.* 21: 67–75.
- Griffiths, R. C., 2006 Coalescent lineage distributions. *Adv. Appl. Probab.* 38: 405–429.
- Griffiths, R. C., and S. Tavaré, 1994a Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B* 344: 403–410.
- Griffiths, R. C., and S. Tavaré, 1994b Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* 46: 131–159.
- Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. *Stoch. Models* 14: 273–295.
- Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5: e1000695.
- Hein, J., M. Schierup, and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, New York.
- Hudson, R. R., 1990 Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7: 44.
- Hudson, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18: 337–338.
- Kimura, M., 1955 Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA* 41: 144–150.
- Kingman, J., 1982a The coalescent. *Stochastic Process. Appl.* 13: 235–248.
- Kingman, J., 1982b Exchangeability and the Evolution of Large Populations, pp. 97–112 in *Exchangeability in Probability and Statistics*, edited by G. Koch and F. Spizzichino. North-Holland, Amsterdam.
- Lukić, S., J. Hey, and K. Chen, 2011 Non-equilibrium allele frequency spectra via spectral methods. *Theor. Popul. Biol.* 79: 203–219.

- Mardis, E. R., 2008 The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24: 133–140.
- Marth, G. T., E. Czubacka, J. Murvai, and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 2004: 351–372.
- Maruvka, Y., N. Shnerb, Y. Bar-Yam, and J. Wakeley, 2011 Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate. *Mol. Biol. Evol.* 28: 1617–1631.
- Nordborg, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop, and C. Cannings. Wiley, Chichester, UK.
- Polanski, A., and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* 165: 427–436.
- Polanski, A., A. Bobrowski, and M. Kimmel, 2003 A note on distributions of times to coalescence, under time-dependent population size. *Theor. Popul. Biol.* 63: 33–40.
- Risch, N., H. Tang, H. Katzenstein, and J. Ekstein, 2003 Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. *Am. J. Hum. Genet.* 72: 812–822.
- Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. *Genetics* 132: 1161–1176.
- Song, Y. S., and M. Steinrücken, 2012 A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics* 190: 1117–1129.
- Tajima, F., 1989 Statistical methods for testing the neutral mutations hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
- Takahata, N., and M. Nei, 1985 Gene genealogy and variance of interpopulation nucleotide differences. *Genetics* 110: 325–344.
- Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* 26: 119–164.
- Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring coalescence times from dna sequence data. *Genetics* 145: 505–518.
- Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7: 256–276.
- Watterson, G., 1984 Lines of descent and the coalescent. *Theor. Popul. Biol.* 26: 77–92.
- Williamson, S. H., R. Hernandez, A. Fedel-Alon, L. Zhu, R. Nielsen *et al.*, 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 A. Rogers, 2002 The matrix coalescent and an application to human single-nucleotide polymorphisms. *Genetics* 161: 1641–1650.
- Živković, D., and W. Stephan, 2011 Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theor. Popul. Biol.* 79: 184–191.

Communicating editor: J. Hermisson

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.151522/-/DC1>

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

Hua Chen and Kun Chen

Table S1 Comparison of the asymptotic approximation and simulated results for the mean and standard deviation of the coalescence time T_m ($N_0 = 2.0 \times 10^6$).

γ	m	Mean (T_m)			Standard Deviation (T_m)		
		Simulation	Asymptotic	Bias (%)	Simulation	Asymptotic	Bias (%)
Sample size n=100							
0.001	5	6607.362	6634.633	27.272 (0.413%)	264.418	270.356	5.937 (2.245%)
0.001	10	5866.462	5888.878	22.416 (0.382%)	199.982	201.996	2.014 (1.007%)
0.001	20	5067.898	5081.404	13.506 (0.267%)	158.140	159.689	1.549 (0.980%)
0.001	50	3700.491	3713.572	13.081 (0.353%)	150.056	149.020	1.036 (0.690%)
0.001	95	1094.259	1133.098	38.839 (3.549%)	294.844	303.323	8.479 (2.876%)
0.005	5	1640.772	1648.604	7.832 (0.477%)	52.380	54.128	1.748 (3.338%)
0.005	10	1494.429	1499.219	4.791 (0.321%)	39.485	40.489	1.004 (2.542%)
0.005	20	1334.856	1337.172	2.316 (0.174%)	31.212	32.097	0.885 (2.835%)
0.005	50	1058.492	1060.661	2.169 (0.205%)	30.216	30.397	0.181 (0.600%)
0.005	95	474.002	488.927	14.924 (3.149%)	82.834	81.717	1.117 (1.349%)
0.010	5	890.541	893.604	3.063 (0.344%)	26.479	27.068	0.589 (2.224%)
0.010	10	816.675	818.897	2.221 (0.272%)	19.925	20.250	0.325 (1.629%)
0.010	20	736.252	737.838	1.586 (0.215%)	16.078	16.059	0.019 (0.119%)
0.010	50	598.054	599.396	1.342 (0.224%)	15.606	15.236	0.370 (2.371%)
0.010	95	299.850	309.343	9.493 (3.166%)	44.908	42.711	2.197 (4.892%)
Sample size n=500							
0.001	5	6644.566	6675.823	31.257 (0.470%)	257.442	259.462	2.020 (0.785%)
0.001	10	5955.294	5973.810	18.515 (0.311%)	183.220	185.641	2.421 (1.321%)
0.001	50	4288.362	4290.459	2.097 (0.049%)	90.942	89.431	1.512 (1.662%)
0.001	100	3494.178	3496.508	2.329 (0.067%)	69.337	69.701	0.363 (0.524%)
0.001	200	2564.372	2564.949	0.577 (0.023%)	61.512	60.764	0.748 (1.216%)
0.001	495	76.400	77.709	1.309 (1.713%)	33.079	33.437	0.358 (1.081%)
0.005	5	1650.639	1656.850	6.211 (0.376%)	50.429	51.945	1.516 (3.006%)
0.005	10	1512.385	1516.242	3.857 (0.255%)	36.212	37.204	0.992 (2.740%)
0.005	50	1177.115	1177.776	0.660 (0.056%)	18.255	18.084	0.171 (0.935%)
0.005	100	1015.878	1016.281	0.403 (0.040%)	14.163	14.286	0.123 (0.869%)
0.005	200	821.763	822.175	0.412 (0.050%)	13.067	12.950	0.117 (0.896%)
0.005	495	66.186	67.871	1.684 (2.545%)	24.898	25.739	0.842 (3.380%)
0.010	5	894.126	897.727	3.602 (0.403%)	25.342	25.976	0.634 (2.502%)
0.010	10	825.872	827.410	1.538 (0.186%)	18.293	18.607	0.313 (1.713%)
0.010	50	657.782	658.064	0.282 (0.043%)	9.002	9.055	0.052 (0.583%)
0.010	100	576.765	577.144	0.379 (0.066%)	7.202	7.165	0.036 (0.501%)
0.010	200	479.310	479.579	0.269 (0.056%)	6.620	6.528	0.091 (1.378%)
0.010	495	57.079	59.227	2.147 (3.762%)	19.395	19.988	0.592 (3.054%)

Table S2 Comparison of the asymptotic approximation and exact results for the mean and standard deviation of $A_n(t)$ (population size $N_0 = 2.0 \times 10^6$).

γ	t	Mean ($A_n(t)$)			Standard Deviation ($A_n(t)$)		
		Exact	Asymptotic	Bias (%)	Exact	Asymptotic	Bias (%)
Sample size n=100							
0.001	100	99.740	99.738	0.003 (0.003%)	0.508	0.511	0.003 (0.497%)
0.001	500	98.420	98.404	0.016 (0.016%)	1.237	1.243	0.006 (0.478%)
0.001	1000	95.921	95.881	0.040 (0.041%)	1.938	1.946	0.009 (0.442%)
0.001	2000	86.347	86.227	0.120 (0.139%)	3.202	3.212	0.010 (0.317%)
0.003	100	99.712	99.709	0.003 (0.003%)	0.535	0.538	0.003 (0.497%)
0.003	500	97.208	97.180	0.027 (0.028%)	1.625	1.632	0.008 (0.460%)
0.003	1000	86.397	86.278	0.119 (0.138%)	3.197	3.208	0.010 (0.317%)
0.003	2000	23.302	22.970	0.332 (1.446%)	2.754	2.750	0.003 (0.127%)
0.010	100	99.577	99.572	0.004 (0.004%)	0.648	0.651	0.003 (0.495%)
0.010	300	95.489	95.446	0.044 (0.046%)	2.029	2.038	0.009 (0.436%)
0.010	600	50.142	49.849	0.293 (0.588%)	3.815	3.815	0.000 (0.011%)
0.010	800	12.173	11.834	0.339 (2.860%)	1.989	1.985	0.005 (0.247%)
Sample size n=400							
0.001	100	395.847	395.837	0.010 (0.003%)	2.017	2.019	0.002 (0.121%)
0.001	500	375.689	375.632	0.057 (0.015%)	4.634	4.639	0.005 (0.103%)
0.001	1000	341.473	341.347	0.126 (0.037%)	6.558	6.563	0.005 (0.076%)
0.001	2000	244.323	244.065	0.258 (0.106%)	7.928	7.929	0.002 (0.020%)
0.003	100	395.400	395.389	0.011 (0.003%)	2.120	2.123	0.003 (0.121%)
0.003	500	358.498	358.405	0.094 (0.026%)	5.785	5.790	0.005 (0.089%)
0.003	1000	244.729	244.471	0.257 (0.105%)	7.928	7.930	0.002 (0.020%)
0.003	2000	28.086	27.750	0.336 (1.210%)	3.044	3.041	0.003 (0.114%)
0.010	100	393.260	393.243	0.017 (0.004%)	2.553	2.556	0.003 (0.119%)
0.010	300	336.029	335.893	0.136 (0.040%)	6.753	6.758	0.005 (0.072%)
0.010	600	79.945	79.613	0.332 (0.416%)	5.133	5.131	0.002 (0.039%)
0.010	800	13.326	12.987	0.339 (2.607%)	2.085	2.081	0.005 (0.230%)

Table S3 Comparison of the gamma approximation and simulated results for the mean and standard deviation of the coalescence time T_m at the early stage of the coalescent (population size $N_0 = 2.0 \times 10^6$).

γ	m	Mean (T_m)			Standard Deviation (T_m)		
		Simulation	Asymptotic	Bias (%)	Simulation	Asymptotic	Bias (%)
Sample size n=500							
0.001	490	149.433	160.321	10.888 (7.286%)	43.256	50.698	7.442 (17.204%)
0.001	494	91.118	96.192	5.074 (5.569%)	35.642	39.270	3.629 (10.181%)
0.001	498	31.402	32.064	0.662 (2.107%)	22.258	22.673	0.415 (1.862%)
0.001	499	15.834	16.032	0.198 (1.253%)	16.259	16.032	0.227 (1.394%)
0.005	490	117.199	160.321	43.121 (36.793%)	28.062	50.698	22.636 (80.663%)
0.005	494	77.306	96.192	18.886 (24.430%)	25.855	39.270	13.415 (51.885%)
0.005	498	28.659	32.064	3.405 (11.881%)	18.154	22.673	4.519 (24.892%)
0.005	499	14.627	16.032	1.405 (9.604%)	13.721	16.032	2.311 (16.845%)
Sample size n=1000							
0.001	990	39.391	40.040	0.649 (1.647%)	12.198	12.662	0.464 (3.800%)
0.001	994	23.675	24.024	0.349 (1.473%)	9.535	9.808	0.273 (2.861%)
0.001	998	7.851	8.008	0.157 (1.997%)	5.519	5.663	0.144 (2.608%)
0.001	999	3.875	4.004	0.129 (3.324%)	4.296	4.004	0.292 (6.795%)
0.005	990	36.280	40.040	3.760 (10.364%)	10.299	12.662	2.363 (22.944%)
0.005	994	22.352	24.024	1.672 (7.480%)	8.907	9.808	0.901 (10.115%)
0.005	998	7.936	8.008	0.072 (0.907%)	5.478	5.663	0.184 (3.362%)
0.005	999	3.992	4.004	0.012 (0.301%)	4.145	4.004	0.141 (3.410%)

Table S4 Comparison of the Poisson approximation and exact results for the mean and standard deviation of coalesced lineage numbers ($n - A_n(t)$) at the early stage of the coalescent (population size $N_0 = 2.0 \times 10^6$).

γ	t	Mean ($n - A_n(t)$)			Standard Deviation ($n - A_n(t)$)		
		Exact	Asymptotic	Bias (%)	Exact	Asymptotic	Bias (%)
Sample size n=500							
0.001	10	0.626	0.627	0.001 (0.125%)	0.790	0.792	0.001 (0.188%)
0.001	20	1.257	1.260	0.003 (0.251%)	1.118	1.123	0.004 (0.377%)
0.001	40	2.533	2.546	0.013 (0.507%)	1.583	1.595	0.012 (0.758%)
0.001	80	5.142	5.195	0.053 (1.028%)	2.244	2.279	0.035 (1.537%)
0.001	100	6.475	6.560	0.085 (1.295%)	2.512	2.561	0.050 (1.933%)
0.005	10	0.639	0.640	0.001 (0.128%)	0.798	0.800	0.002 (0.192%)
0.005	20	1.309	1.312	0.003 (0.262%)	1.141	1.145	0.004 (0.392%)
0.005	40	2.747	2.762	0.015 (0.549%)	1.648	1.662	0.014 (0.822%)
0.005	80	6.061	6.136	0.074 (1.212%)	2.432	2.477	0.045 (1.810%)
0.005	100	7.964	8.093	0.129 (1.593%)	2.777	2.845	0.068 (2.375%)
Sample size n=1000							
0.001	10	2.504	2.510	0.006 (0.250%)	1.578	1.584	0.006 (0.375%)
0.001	20	5.020	5.045	0.025 (0.502%)	2.229	2.246	0.017 (0.752%)
0.001	40	10.090	10.192	0.103 (1.009%)	3.144	3.193	0.048 (1.508%)
0.001	80	20.377	20.801	0.424 (2.038%)	4.422	4.561	0.138 (3.034%)
0.001	100	25.594	26.266	0.672 (2.559%)	4.930	5.125	0.195 (3.804%)
0.005	10	2.554	2.561	0.007 (0.255%)	1.594	1.600	0.006 (0.383%)
0.005	20	5.226	5.253	0.027 (0.523%)	2.274	2.292	0.018 (0.782%)
0.005	40	10.938	11.059	0.121 (1.094%)	3.271	3.326	0.054 (1.634%)
0.005	80	23.978	24.567	0.589 (2.398%)	4.780	4.956	0.177 (3.566%)
0.005	100	31.387	32.404	1.017 (3.139%)	5.427	5.692	0.265 (4.655%)

Figure S1 The gamma distribution as the asymptotic distribution of coalescence times at the early stage of the coalescent.

(A), (C) and (E): The asymptotic probability density functions of coalescence times in a sample of size 1000, collected from a population with $N_0 = 2 \times 10^6$. The growth rate is $\gamma = 0.001$. The plots from top to bottom correspond to the p.d.f.'s of T_{998} , T_{994} and T_{990} . The histograms were generated by 500 coalescent simulations.

(B), (D) and (F): The same settings as (A), (C) and (E), except that $\gamma = 0.005$.

Figure S2 The Poisson distribution as the asymptotic distribution of coalesced lineages ($n - A_n(t)$) at the early stage of the coalescent.

(A), (C) and (E): The asymptotic distribution of coalesced lineages at different times in a sample of size 1000, collected from a population with $N_0 = 2 \times 10^6$. The growth rate is $\gamma = 0.001$. The plots from top to bottom correspond to the distribution at time $t = 10$, $t = 40$ and $t = 80$. The histograms were generated by 500 coalescent simulations.

(B), (D) and (F): The same settings as (A), (C) and (E), except that $\gamma = 0.005$.