Inferring Very Recent Population Growth Rate from Population-Scale Sequencing Data: Using a Large-Sample Coalescent Estimator

Hua Chen,*[,1] Jody Hey[2] and Kun Chen[3]

[1]Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[2]Center for Computational Genetics and Genomics, Temple University

[3]Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA

***Corresponding author:** E-mail: chenh@big.ac.cn.

Associate editor: Matthew Hahn

Abstract

Large-sample or population-level sequencing data provide unprecedented opportunities for inferring detailed population histories, especially recent demographic histories. On the other hand, it challenges most existing population genetic methods: Simulation-based approaches require intensive computation, and analytical approaches are often numerically intractable when the sample size is large. We propose a computationally efficient method for simultaneous estimation of population size, the rate, and onset time of population growth in the very recent history, using the pattern of the total number of segregating sites as a function of sample size. Coalescent simulation shows that it can accurately and efficiently estimate the parameters of recent population growth from large-scale data. This approach has the flexibility to model population history with multiple growth stages or other epochs, and it is robust when the sample size is very large or at the population scale, for which the Kingman's coalescent assumption is not valid. This approach is applied to recently published data and estimates the recent population growth rate in the European population to be 1.49% with the onset time 7.26 ka, and the rate in the African population to be 0.735% with the onset time 10.01 ka.

Key words: **genetics diversity, population growth rate, large-sample sequencing, coalescent.**

Introduction

Traditional population genetic inference methods were developed when large-scale genomic polymorphism data were scarce. These methods typically work for a small sample (e.g., less than 100 haplotypes) and focus more on parameter inference of relatively ancient events. Population genetic studies that are based on allele frequency spectrum (AFS) or linkage disequilibrium at genomic level in a small or moderate sample have reported reasonably tight confidence intervals for parameters of relatively ancient events within the time range of 20 ky–3 My (Li and Durbin 2011; see also Schaffner et al. 2005; Gutenkunst et al. 2009; Gravel et al. 2011; Lukić and Hey 2012). In contrast, the inference of very recent demographic events relies heavily on observing mutations that occurred very recently, which are rare and only detectable in large samples.

With the reduction of sequencing cost, whole-genome sequence data of a large number of individuals that account for a significant portion of entire population are becoming common, especially in disease studies (Altshuler et al. 2010; Coventry et al. 2010; Nelson et al. 2012; Tennessen et al. 2012; Fu et al. 2013). Large-scale genetic polymorphism data encourage the exploration of fine-scale demographic events that happened in recent history. Inferring the recent demographic history is of great interests in both population genetic studies and disease studies, as recent demographic history has a significant effect on shaping the genetic variation in modern human populations, and understanding the interaction between demographic and genetic factors helps the design of studies on inherited diseases with different underlying genetic architecture (Gravel et al. 2011). In a recent study of two candidate gene regions, more than 13,000 individuals were sequenced (Coventry et al. 2010). The authors demonstrated that single nucleotide polymorphisms (SNPs) discovered in such a large sample were consistent with a very recent and rapid population growth model. Gravel et al. (2011) investigated the sequencing data in the pilot phase of the Thousand Genomes Project and found an excessive number of rare and population-specific mutants with increased sampling of individuals. Nelson et al. (2012) analyzed exome sequencing data from 202 genes in 14,000 individuals, and Tennessen et al. (2012) analyzed 15,585 genes in 2,440 individuals with European and African ancestry. Both studies reported similar findings.

A useful piece of information that can be used to infer very recent population size and growth rate is the total number of segregating sites as a function of sample size (TNSFS). The TNS is expected to increase with sample size, and the dynamics of TNSFS trajectory provides information for parameter inference, as was first proposed by Coventry et al. (2010). They fit an exponential population growth model to the TNSFS trajectory and AFSs of two genes, and estimated the growth rate in the European population to be 0.094. Nelson et al. (2012) did similar analysis with their data and estimated the

doi:10.1093/molbev/msv158 Advance Access publication July 16, 2015

recent growth rate of Europeans to be 0.017. To infer recent population size and growth rate using TNSFS, Coventry et al. (2010) and Nelson et al. (2012) used forward and coalescent simulations, respectively, to generate random samples for a range of population sizes and growth rates, and estimated the parameters of interest from the simulated data that were compatible with the TNSFS or AFS observed from the real data. Both of their approaches require intensive simulations to obtain accurate and reliable estimates, and the computation becomes extremely time consuming when the sample size is large.

To gain computational efficiency, a method matching the observed TNSFS to the theoretical prediction in analytical form will be preferable to those based on simulation. However, when sample size n is large, direct application of existing analytical formula for TNS faces three major challenges. First, the existing formulas have numerical issues for large n. For example, the terms of the alternating-sum series in the exact formula of AFS are exploding when $n > 100$ (Polanski and Kimmel 2003). Second, the existing formulas are for the simple exponential growth model, and it may be nontrivial to extend it to more complex demographic models, such as, a multiple-stage exponential growth model. Third, when a large proportion of the entire population is sampled, one of the assumptions in Kingman's coalescent, which the sample size is far less than the population size ($n \ll N$), is violated, so that multiple lineages can coalesce in a single generation during the exact coalescent process. In this case, the Kingman's coalescent may have serious deviation from the exact coalescent process under the Wright–Fisher model, and the validity of the conventional inference methods based on the Kingman's coalescent needs to be assessed for large samples. For the above reasons, simulations, especially population-level forward simulations, were often adopted to generate the exact coalescent process for large samples (Coventry et al. 2010; Nelson et al. 2012).

New Approaches

The method we propose is based on the pattern of the TNSFS. To set up the method, we derive the analytical formula for the expected TNS for populations underwent one, or multiple growth stages. We fit the expected TNSFS to the observed TNSFS trajectory by nonlinear least squares (NLS) techniques, to make inference on the demographic parameters.

Being different from Coventry et al. (2010) and Nelson et al. (2012), the formula for TNS is analytical and in simple form, which guarantees computational efficiency. The formula is derived based on asymptotic theory of coalescent distributions, and does not suffer from the numerical issues caused by large sample size (Chen H and Chen K 2013). The NLS fitting method can be implemented promptly. We use simulation to demonstrate the accuracy and reliability of the proposed method for a wide range of parameter values. Another reason we choose the TNSFS is because of its robustness under the violation of the Kingman's coalescent assumption, when the sample size is at the population scale and is a large portion of the population. Through simulation studies, it is demonstrated that for a wide range of sample size TNSFS has less deviation from Kingman's theories than the AFS, suggesting the proposed method to be a very useful solution for population level sequencing data. The article is closed by the application of our proposed method to two recently published data.

A Simple Exponential Growth Model

To set up the method, we first show the exact formula of the expected TNS for a simple exponential growth model, which includes two parameters: The contemporary population size N and the exponential growth rate r, and then derive the asymptotic expectation of the TNS under the simple exponential growth model for large n.

Exact Formula for the Total Number of Segregating Sites

Watterson's θ_W, as a classic measure of genetic diversity, is related to the TNS of a sample in a stationary population by:

$$\mathbb{E}S_W = \theta_W \times \sum_{i=1}^{n-1} \frac{1}{i}, \tag{1}$$

where $\mathbb{E}S_W$ is the expected TNS that can be identified from a sample of n haplotypes (Watterson 1975). Watterson's θ_W estimated from equation (1) can be used as an unbiased estimator of the scaled mutation rate $2N\mu$ in populations with constant size: If mutations are assumed to follow an infinitely many-sites model and occur along branches of the gene genealogy following a Poisson process, the expected TNS of a sample can be estimated by the product of the mutation rate, μ, and the expected total branch length (ETBL) of the gene genealogy, which is $2N\sum_{i=1}^{n-1}\frac{1}{i}$ for a haploid population of constant size N (Hudson 1990; Fu 1995). For a population at nonequilibrium, or a population undergoing expansion or contraction, θ_W can still be estimated from equation (1), but it is no longer an unbiased estimator of the scaled mutation rate, as $2N\sum_{i=1}^{n-1}\frac{1}{i}$ is not an unbiased estimator of the ETBL. As illustrated in figure 2 of Coventry et al. (2010), the extrapolation of $\mathbb{E}S_W$ by plugging Watterson's θ_W into equation (1) for larger n predicted far less segregating sites than what can be actually observed in larger samples. This was explained in Coventry et al. (2010) by the rapid population growth in recent human history. If we can explicitly derive the ETBL of gene genealogies under any demographic model, an analog of $\mathbb{E}S_W$ in a nonequilibrium population can be written as:

$$\mathbb{E}S = \mu \cdot \mathbb{E}\mathrm{TBL} = \mu \cdot \sum_{m=2}^{n} m\mathbb{E}W_m, \tag{2}$$

where $\mathbb{E}W_m$ denotes the expected intercoalescence time during which there are m lineages in the genealogy.

The exact marginal distribution and expectation of intercoalescence times in a population with temporally varying size were derived by Polanski et al. (2003) using the method of integral transform. In Appendix A, we show that $\mathbb{E}W_m$ can be obtained with a simple approach, and in specific, for a population under the simple exponential growth model:

$$\mathbb{E}W_m = \sum_{i=m}^{n} A_i^{n,m} e^{\frac{i(i-1)}{2Nr}} \frac{1}{r} \mathrm{Ei}\left(-\frac{i(i-1)}{2Nr}\right), \tag{3}$$

where Ei $(\cdot)$ stands for the exponential integral, and $A_i^{n,m}$ is the coefficient of the alternating hypergeometric series (eq. A2).

The estimation of $\mathbb{E}W_m$ using equation (3) for large-sample genealogies is not trivial, and the value of individual term in the hypergeometric series explodes with the increase of sample size n, which limits its practical application to small samples (e.g., $n < 100$). But the evaluation of ETBL, as pointed out by Polanski and Kimmel (2003), can be simplified by interchange of the double summation (see eq. 5), so that avoid the calculation of coefficient terms ($A_i^{n,m}$) with potential memory overflow for large n. Then, the resulting equation can be applied to quite large samples:

$$\begin{aligned} \mathbb{E}S &= \sum_{m=2}^{n} kW_m \\ &= \sum_{m=2}^{n}\sum_{i=m}^{n} A_i^{n,m} e^{\frac{i(i-1)}{2Nr}} \frac{1}{r} \mathrm{Ei}\left(-\frac{i(i-1)}{2Nr}\right) \end{aligned} \tag{4}$$

$$= \mu \sum_{i=2}^{n} \frac{1}{r} e^{\frac{i(i-1)}{2Nr}} \mathrm{Ei}\left(-\frac{2i(2i-1)}{2Nr}\right) V_i^n, \tag{5}$$

with $V_i^n = (2i-1)\frac{n!(n-1)!}{(n+i-1)!(n-i)!}[1+(-1)^i]$. In figure 1, the expected TNS based on equation (5) was plotted as a function of the sample size for different growth rates, $r = 0.001, 0.002, 0.005, 0.01,$ and 0.02, with the mutation rate $\mu = 1 \times 10^{-8}$ per site per generation for a 100-kb region.

Asymptotic Formula of TNS for Large-Sample Genealogies

One approach to avoid numerical instability in the exact formula is to approximate the exact expectation of TNS by the asymptotic one. The asymptotic expectation of TBL can be derived as follows:

$$\begin{aligned} \mathbb{E}\mathrm{TBL}^{asy} &= \mathbb{E}\left[\int_0^{T_1} \mathbb{E}A_n(t)\,\mathrm{d}t \mid T_1\right] \\ &\approx \mathbb{E}_{T_1}\left[\int_0^{T_1} \frac{n}{(1-\frac{n}{2N_0r}) + \frac{n}{2N_0r}e^{rt}}\mathrm{d}t\right] \\ &\approx \mathbb{E}_{T_1}\left[\frac{2nN_0\log[(1-e^{-rT_1})\frac{n}{2N_0r} + e^{-rT_1}]}{n-2N_0r}\right] \\ &\approx \mathbb{E}_{T_1}\left[\frac{2nN_0\log(n/2N_0r)}{n-2N_0r}\right] \\ &= \frac{2nN_0\log(n/2N_0r)}{n-2N_0r}, \end{aligned} \tag{6}$$

where T_i is the coalescence time when $i+1$ lineages coalesce into i lineages, and $A_n(t)$ is the number of ancestral lineages of the contemporary n haplotypes at time t. The approximation

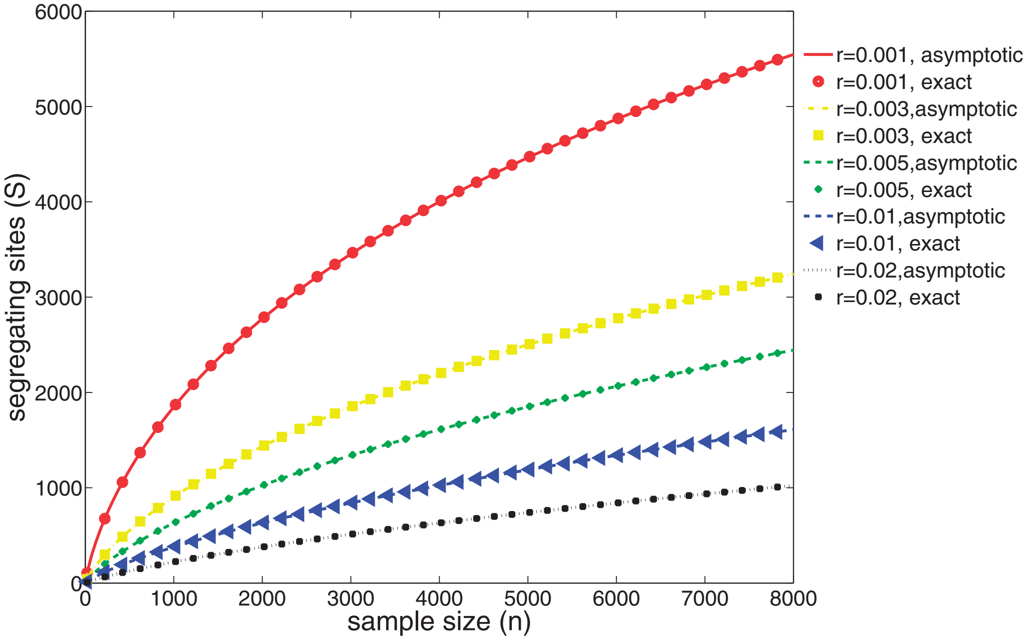

Fig. 1. The expected TNS ($\mathbb{E}S$) as a function of haploid sample size (n). The points represent the exact expectation given by equation (5), and the lines represent asymptotic approximation (eq. 7). Different types of lines and points represent different population growth rates: $r = 0.001, 0.002, 0.005, 0.01,$ and 0.02. The population size is $N_0 = 2 \times 10^6$, the mutation rate is chosen to be 1×10^{-8} per site per generation, and the whole region spans 100 kb.

in the second line holds as the asymptotic expectation of $A_n(t)$ for the exponential growth model is shown to be $\frac{n}{(1-\frac{n}{2N_0r})+\frac{n}{2N_0r}e^{rt}}$ in Chen H and Chen K (2013), and the fourth line holds as rT_1 is quite large so that e^{-rT_1} is ignorable. Consequently, we have

$$\mathbb{E}S = \mu \cdot \mathbb{E}\mathrm{TBL}^{asy} = \frac{2nN_0\mu\log(n/2N_0r)}{n-2N_0r}. \tag{7}$$

The above equation can be used to define an analog of Watterson's diversity measure for populations under simple exponential growth:

$$\theta_g \equiv 2N_0\mu = \frac{\mathbb{E}S}{\frac{n\log(n/2N_0r)}{n-2N_0r}}. \tag{8}$$

We can now describe the expected TNS as a function of the sample size for exponentially growing populations. In figure 1, the expected TNSs given by equation (7) for different growth rates were plotted together with the exact expectation given by equation (5). The asymptotic expectation of the TNS approximates the exact expectation very well, and works for the sample sizes that have been tested here, ranging from 20 to 8,000. Compared with the exact equation, the asymptotic equation is in simpler form, its computation is very efficient, and evaluating the formula does not require handling numerical issues even for ultra-large sample sizes.

Two-Stage Exponential Growth Model

The two-stage exponential growth (TEG) model, which assumes a stage of constant population size followed by a stage of exponential growth, was more commonly used in the population genetic inference (Adams and Hudson 2004; Chen et al. 2007; Evans et al. 2007; Coventry et al. 2010; Chen 2013). The TEG model has three parameters: The onset time of population growth τ, the exponential growth rate r, and the contemporary population size N_0. The ancestral population size at time τ, N_a, can also be treated as a free parameter. In such a case, a jump in the population size at time τ is allowed. The two-stage model for population size can be written as:

$$N(t) = \begin{cases} N_0e^{-rt}, t \le \tau \\ N_a, t > \tau. \end{cases}. \tag{9}$$

In the derivation of the asymptotic properties of coalescence times and number of ancestral lineages, we often use a scaling function of time t, $g(t) = \int_0^t \frac{1}{N(u)}du$ (Griffiths and Tavaré 1994, 1998). Specifically for the exponential growth stage with rate r, we have $g(t) = (e^{rt}-1)/(N_0r)$. The time-scaling function $g(t)$ is important for deriving the ETBL for different population growth models, as it is the essential component in the asymptotic distribution of $A_n(t)$. It was shown that (Chen H and Chen K 2013),

$$\mathbb{E}A_n(t) = u_t \approx \frac{n}{1+\frac{n}{2}g(t)},$$

and

$$\mathrm{Var}(A_n(t)) = \sigma_t^2 \approx \frac{n(1-(1+\frac{n}{2}g(t))^{-3})}{3(1+\frac{n}{2}g(t))}.$$

The TBL for the TEG model can be written as the sum of two parts: The TBL in the growth phase ($\mathrm{TBL}^g(n)$) and the TBL in the constant phase ($\mathrm{TBL}^a(n)$), or $\mathbb{E}\mathrm{TBL}(n) = \mathbb{E}\mathrm{TBL}^a(n) + \mathbb{E}\mathrm{TBL}^g(n)$. With more details shown in Appendix B, we have

$$\begin{aligned} \mathbb{E}(\mathrm{TBL}^g(n)) &= \int_0^\tau \mathbb{E}A_n(t)dt \\ &\approx -\frac{2nN_0\log[(1-e^{-r\tau})\frac{n}{2N_0r}+e^{-r\tau}]}{2N_0r-n}, \end{aligned} \tag{10}$$

and $\mathbb{E}\mathrm{TBL}^a(n)$ can be obtained by taking expectation of $\mathbb{E}(\mathrm{TBL}^a(n) \mid A_n(\tau)=m)$ with respect to m, where $A_n(\tau)$ is the number of ancestral lineages at time τ:

$$\begin{aligned} \mathbb{E}(\mathrm{TBL}^a(n)) &= \mathbb{E}(\mathbb{E}(\mathrm{TBL}^a(n) \mid A_n(\tau)=m)) \\ &= 2N_a\mathbb{E}\left(\sum_{i=1}^{m-1}\frac{1}{i}\right) \approx 2N_a\mathbb{E}(\log(m)+\gamma) \\ &\approx 2N_a\left(\log(u_\tau)+\gamma-\frac{\sigma_\tau^2}{2u_\tau^2})\right) \\ &\approx 2N_a\left(\log\left(\frac{n}{1+\frac{n}{2}g(\tau)}\right)+\gamma\right), \end{aligned} \tag{11}$$

with $\gamma = 0.57721566$ being the Euler constant; and u_τ and σ_τ being the mean and standard deviation of $A_n(\tau)$, respectively. The approximation in the second line of equation (11) is accurate when m is large (Watterson 1975). The remaining term $\frac{\sigma_\tau^2}{2u_\tau^2}$ in the third line can be ignored as σ_τ^2 is relatively very small when compared with u_τ (Chen H and Chen K 2013). Combining the two parts of ETBL, we have

$$\begin{aligned} \mathbb{E}(\mathrm{TBL}(n)) &= \mathbb{E}(\mathrm{TBL}^g(n)) + \mathbb{E}(\mathrm{TBL}^a(n)) \\ &= -\frac{2nN_0\log[(1-e^{-r\tau})\frac{n}{2N_0r}+e^{-r\tau})]}{2N_0r-n} \\ &\quad + 2N_a\left(\log\left(\frac{n}{1+\frac{n}{2}g(\tau)}\right)+\gamma\right). \end{aligned} \tag{12}$$

The expected TNS can then be estimated by taking the product of ETBL and mutation rate μ. In figure 2, we presented the theoretical results based on equation (12) for different growth rates: $r = 0.001, 0.003, 0.005, 0.01,$ and 0.02, as well as coalescent simulated $\mathbb{E}S$. The theoretical result fits the simulation very well for a wide range of growth rates and sample sizes.

Extension to Multiple Stages

More complicated multistage growth models were also proposed. For example, Keinan and Clark (2012) proposed a multiepoch exponential growth model, in which one epoch of moderate exponential growth is followed by an explosive

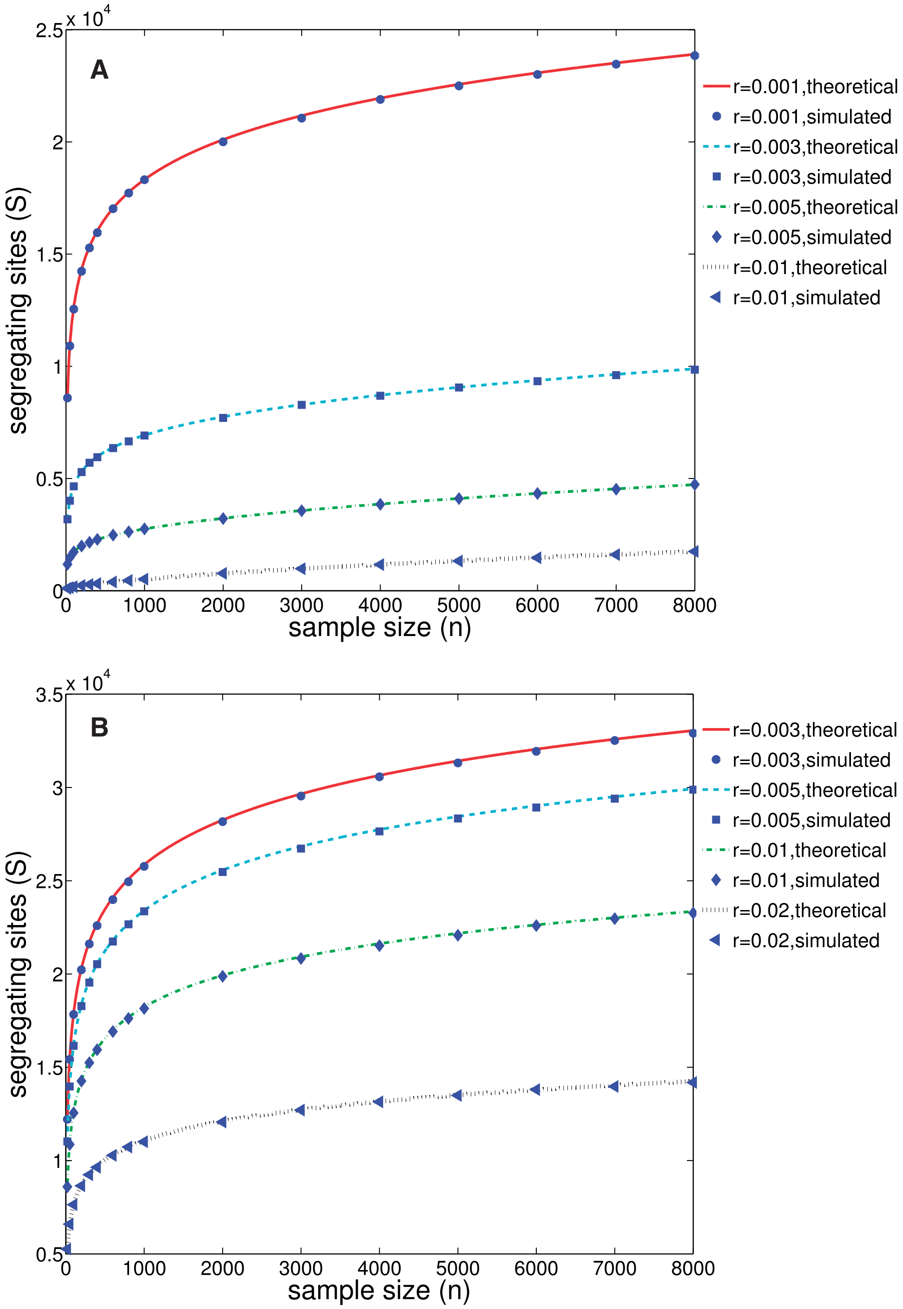

FIG. 2. The theoretical prediction, represented by lines, and simulated results, represented by points, of TNSFS in different parameter settings under a TEG model. (*A*) The TNSFS under a two-stage model with the initial time of growth phase $\tau = 500$ generations and the contemporary population size $N_0 = 2 \times 10^6$. The different curves correspond to theoretical predictions for different growth rates: $r = 0.001, 0.003, 0.005$, and 0.01. The sample sizes of the simulated data are set to be $n = 20, 50, 100, 200, 300, 400, 600, 800, 1,000, 2,000, 3,000, 4,000$, 5,000, 6,000, 7,000, and 8,000. (*B*) The TNSFS under a two-stage model with the initial time of growth phase $\tau = 50$ generations and the contemporary population size $N_0 = 2 \times 10^6$. The growth rates are chosen to be $r = 0.003, 0.005, 0.01$, and 0.02. The other parameters are set to be the same as (*A*).

growth. If we make a little modification of Keinan and Clark's model, a three-stage exponential growth model can be written as:

$$N(t) = \begin{cases} N_0 e^{-r_1 t}, t \le \tau_1 \\ N_0 e^{-r_1\tau_1 - r_2(t-\tau_1)}, \tau_1 \le t \le \tau_2 \\ N_a, t > \tau_2. \end{cases} \quad (13)$$

The time-scaling function $g(t)$ can be achieved from the above population growth model:

$$g(t) = \begin{cases} \frac{1}{N_0 r_1}(e^{r_1 t} - 1), t \le \tau_1, \\ \frac{e^{r_1\tau_1}}{N_0 r_2}(e^{r_2(t-\tau_1)} - 1) + \frac{1}{N_0 r_1}(e^{r_1\tau_1} - 1), \tau_1 \le t \le \tau_2. \end{cases} \quad (14)$$

Using $g(t)$, the ETBL for the two growth stages can be similarly obtained as in previous sections (eq. 15, see Appendix C for more details).

$$\begin{aligned}\mathbb{E}\mathrm{TBL}^9 = {} & \frac{-2nN_0}{2N_0r_1-n}log\left[(1-e^{-r_1\tau_1})\frac{n}{2N_0r_1}+e^{-r_1\tau_1}\right]\\ & -\frac{2nN_0r_1}{ne^{r_1\tau_1}(r_2-r_1)-nr_2+2N_0r_1r_2}\\ & \times\{log[r_2(2nN_0r_1-n)+nr_1e^{r_1\tau_1-r_2\tau_1+r_2\tau_2}\\ & +n(r_2-r_{1)}e^{r_1\tau_1}]-log[r_2(n(e^{r_1\tau_1}-1)+2N_0r_1)]\\ & +r_2\tau_1-r_2\tau_2\}.\end{aligned}\tag{15}$$

Gazave et al. (2014) proposed a more complex model for the European demography, which includes a first stage of exponential growth, and then followed by five durations with different population sizes. The piecewise function of the population size is:

$$N(t)=\begin{cases}N_1e^{-rt}, t\le\tau_1,\\ N_k, \tau_{k-1}<t\le\tau_k, 2\le k\le M,\end{cases}\tag{16}$$

with $\tau_M=\infty$, M = 6. The ETBL of the Gazave model is:

$$\mathbb{E}\mathrm{TBL}=\mathbb{E}\mathrm{TBL}_1+\sum_{k=2}^{5}\mathbb{E}\mathrm{TBL}_k+\mathbb{E}\mathrm{TBL}_6,\tag{17}$$

and the equation for each ETBL_k can be found in Appendix D.

In figure 3, we presented the theoretical results of the Gazave model for different growth rates: $r=0.005, 0.01, 0.03,$ and 0.05, together with $\mathbb{E}S$ from coalescent simulation. The two results fit well again for the tested parameter values. Thus, our method is flexible to accommodate various complicated and realistic demographic histories.

Approximation for Exact Coalescent in Ultra-Large or Population-Level Samples

We have so far derived the asymptotic expectation of large-sample TNS for simple, two-, three-, and multiple-stage exponential growth models (eqs. 7, 12, 15, and 17). All the theoretical work is constructed under the Kingman's coalescent framework, as we have assumed that the population size is large and the sample size is much smaller than the population size even for large gene genealogies. With this assumption, Wright–Fisher model finds nearly all probability masses jumping from i lineages to i or $i-1$ ancestors in the immediate previous generation, so that the Wright–Fisher model is well approximated by Kingman's coalescent. However when the sample size is ultra large that even accounts for a significant proportion of the population, the assumption of $n << N$ does not hold anymore. When n is close to N, there could be multiple collisions of lineages during a single generation (Wakeley and Takahashi 2003; Fu 2006; Bhaskar et al. 2014). Such a coalescent process is called "exact coalescent" (Wakeley and Takahashi 2003; Fu 2006). Whether the theoretical results derived based on Kingman's coalescent can be directly used for the exact coalescent has been discussed in former theoretical studies for constant populations (Fu 2006). Both simulation and theoretical studies demonstrated that in constant populations, some characteristics of Kingman's coalescent, such as the TBL of a genealogy, provide remarkably accurate approximation to the exact coalescent (Fu 2006), whereas some other characteristics, such as the AFS, are less accurate. The validity of this conclusion was yet to be addressed in populations with temporally varying size.

In this section, we performed simulation studies for both the Kingman's coalescent and exact coalescent processes, and compared the ETBL and AFS estimated from the simulated data to assess the performance of Kingman's coalescent in approximation of the exact coalescent. For the comparison of AFS, we focus on singletons as in Bhaskar et al. (2014), and use EL_1, the expected external branches of gene genealogies to represent the number of singletons. We estimated ETBL and EL_1 as the average of 2,000 generated samples under Kingman's coalescent with Hudson's simulator "ms," and obtained ETBL and EL_1 of exact coalescent samples with a code modified from Bhaskar et al. (2014). The details of the exact coalescent can be found in the original literature. We compare ETBL and EL_1 for two models: The modified Gazave model (MGM) and the TEG model. The MGM was used by Bhaskar et al. (2014), in which the recent growth phase is approximated by a constant haploid population size of 20, 170, and all the other phases, including the two bottlenecks are identical to the original Gazave model. The TEG model was simulated with a current population size $N_0=20{,}000$, and three growth rates: $r=0.005, 0.01,$ and 0.05. For each simulation, we ranged the sample sizes from 500 to 18,000. The relative bias was calculated as $\frac{|\mathrm{ETBL}_{exact}-\mathrm{ETBL}_{Kingman}|}{\mathrm{ETBL}_{exact}}$ and shown in table 1. As we can see from table 1, ETBLs of the two coalescent processes are quite close for different models and various growth rates. Even when a large proportion of the current population is sampled, the relative bias is around 2%. AFS is less robust to the violation of Kingman's coalescent assumption. The relative bias of EL_1 remains small even when the sample size is large, but it becomes significant when the sample is a large proportion of the current population (e.g., $n>$ 10,000). The relative bias of ETBL and EL_1 increasing with sample size shows similar trend cross different models. This observation is consistent with former studies (Fu 2006; Bhaskar et al. 2014). Based on our simulations and Fu (2006)'s work, our equations of ETBL (eqs. 7, 12, 15, and 17) derived in previous sections using Kingman's coalescent theory are still good estimators of ETBL when the sample size is a significant proportion of or close to the population size. This is especially useful for samples from medical studies of small populations, such as the founder or isolated populations. Note that the population size we chose in table 1 is relatively small (20,000), but as pointed out by Fu (2006), the relative bias of the ETBL is largely a function of n/N instead of absolute N, the property presented in table 1 is thus valid for other population sizes.

An NLS Fitting Method for Inferring Population Growth Rate and Time

Parameter Inference

It has been shown in Coventry et al. (2010) that the upward trend of the TNS with the increase of sample size provides

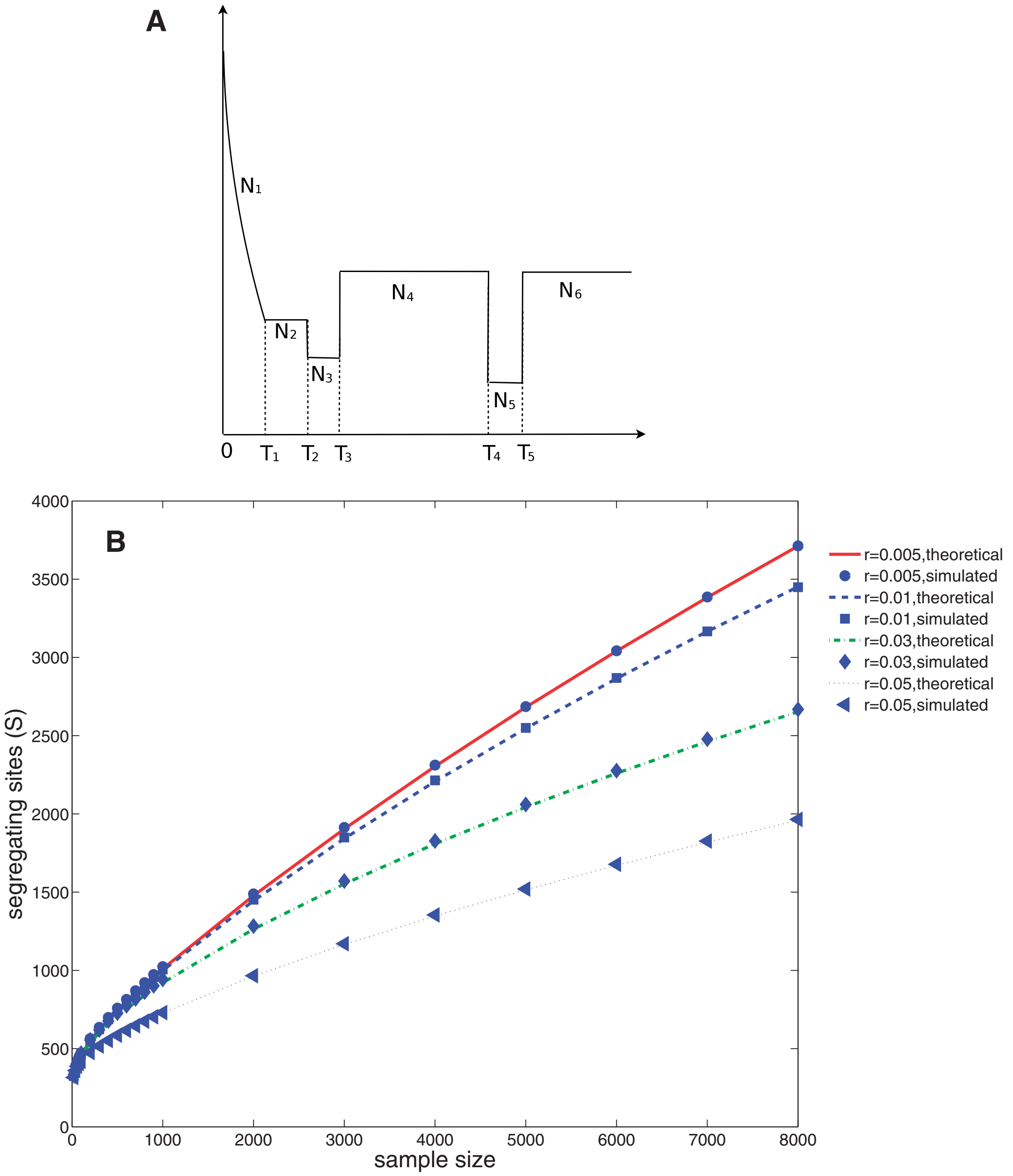

Fig. 3. The theoretical prediction, represented by lines, and simulated results, represented by points, of TNSFS in different parameter settings under the Gazave model. (*A*) Illustration of the Gazave model. (*B*) The TNSFS under the Gazave model. The growth rates are chosen to be $r = 0.005, 0.01, 0.03$, and 0.05. The other parameters are set to be the same as the original Gazave model.

Table 1. The Relative Bias of $\mathbb{E}$TBL and $\mathbb{E}L_1$ for Different Population History Models, When Using the Kingman's Coalescent to Approximate the Exact Coalescent.

Population Model	Statistics	Sample Size					
		500	1,000	2,000	10,000	15,000	18,000
Gazave Model	TBL	0.38%	0.15%	0.25%	0.38%	0.11%	1.18%
	L_1	0.30%	0.42%	0.95%	4.80%	7.53%	9.37%
TEG, $r = 0.005$	TBL	0.066%	0.064%	0.20%	0.85%	1.46%	1.37%
	L_1	0.30%	0.50%	1.12%	5.26%	8.21%	10.02%
TEG, $r = 0.01$	TBL	0.36%	0.81%	0.61%	1.25%	1.86%	1.95%
	L_1	0.44%	0.54%	1.06%	5.56%	8.57%	10.53%
TEG, $r = 0.05$	TBL	0.21%	0.33%	0.56%	1.64%	2.01%	2.52%
	L_1	0.27%	0.61%	1.49%	6.83%	10.21%	12.37%

useful information for parameter inference on the rate and the onset time of recent population growth. We name such a data pattern by "the TNSFS." As seen from our derived asymptotic expectation of TNS for different growth models, the TNS is a nonlinear function of the sample size n, population size N_0, population growth onset time τ, and growth rate r. We therefore fit the expected TNSFS to the observed trajectory with NLS for estimation of parameters involved. The observed TNS for a fixed subsample size was obtained by averaging the TNS in a number of subsamples of the given size that is randomly drawn from the original sample. With that in mind, an observed trajectory of TNSFS in this context is the averaged TNSFS curve.

To be more specific, the data used in our method consist of a pair of vectors: The subsample size, $x(i)$, and the corresponding average count of TNS, $S(x(i))$, for $2 \leq x(1) \leq \cdots \leq x(n-1) \leq n$. Ideally, we can obtain all $n-1$ data points. Sometimes, for the convenience, only d of the $n-1$ points are sampled. We denote the simplified data by $D = \{(x(i), S(x(i))), 2 \leq x(i) \leq n, 1 \leq i \leq d\}$. Now we focus on the parameter inference of a TEG model with no population size jump at τ, and other growth models can be implemented similarly. The parameters of this model include the growth rate r, the onset time of the growth phase τ, and the present population size N_0. Given the parameters, the expected TNS, $\mathbb{E}S(i) = f(x(i) \mid r, \tau, N_0)$, can be analytically estimated through the formulae derived in previous sections (eq. 12). The optimization function aims to minimize the squared error over the parameters:

$$\underset{r,\tau,N_0}{\text{minimize}} \sum_{i=1}^{d} \Big(f(x(i) \mid r, \tau, N_0) - S(i)\Big)^2. \qquad (18)$$

The Levenberg–Marquardt algorithm was used here to obtain the least-squares fitting estimates (Galassi et al. 2009).

Multiple Loci with Heterogeneous Mutation Rates

In the above section, we assumed that the mutation rates are uniform across the genome so that the segregating sites discovered in different regions can be merged together to calculate the TNSFS and infer parameters by the NLS. However, it is well known that high heterogeneity in mutation rates exists across the genome (Tyekucheva et al. 2008). If the local mutation rates are known from other sources, the optimization function in equation (18) can be modified to account for the influence of heterogeneity on the TNS:

$$\underset{r,\tau,N_0}{\text{minimize}} \sum_{l=1}^{L} \sum_{i=1}^{d} \Big(f(x(i,l) \mid r, \tau, N_0, \mu_l) - S(i,l)\Big)^2, \qquad (19)$$

where L is the number of loci, and μ_l is the local mutation rate for locus l and assumed to be known.

If the local mutation rates are unknown, we adopt two approaches to model the mutation rate heterogeneity. In the first approach, we introduce a new parameter for each locus, $\lambda_l = \frac{\mu_l}{\mu_0}$, the relative mutation rate of locus l over the genome averaged mutation rate μ_0. In the second approach, we assume the relative mutation rates from different loci follow a gamma distribution (Nei et al. 1976; Wakeley 1993; Yang 1994):

$$h(\lambda \mid \alpha) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\alpha\lambda}, \qquad (20)$$

where α is the shape parameter. The above distribution is a specific case of gamma distribution in which the rate parameter equals the shape parameters $\beta = \alpha$, and thus $\mathbb{E}\lambda = 1$ and $\text{Var}(\lambda) = 1/\alpha$. As it is difficult to obtain the integral of $f(x(i) \mid r, \tau, N_0, \alpha)$ with respect to the gamma distribution, we use the discrete approximation of the integral following Yang (1994)'s scheme. We discretize the p.d.f. of the gamma distribution into k categories with equal probability in each category $p(\lambda_j \mid \alpha) = 1/k$ and λ_j being the mean of each category. Summing over the probabilities of the k categories, the optimization function under heterogeneous mutation rate now becomes:

$$\underset{r,\tau,N_0,\alpha}{\text{minimize}} \sum_{l=1}^{L} \sum_{i=1}^{d} \Big(f(x(i) \mid r, \tau, N_0, \alpha) - S(i)\Big)^2. \qquad (21)$$

with $f(x(i) \mid r, \tau, N_0, \alpha) = \sum_{j=1}^{k} f(x(i) \mid r, \tau, N_0, \lambda_j) \times p(\lambda_j \mid \alpha)$. Note that in both approaches, we do not estimate the averaged mutation rates across different loci, μ_0, because in practice, we found it is not very informative to estimate the absolute mutation rates jointly with the other population genetic parameters.

Confidence Intervals for Parameters

We used the resampling method to compute the confidence intervals for the estimated parameters (Efron 1982). Fixing the parameters, such as onset time and rate of the growth phase, at the values estimated from above procedure, we can generate a sample of n haplotypes using coalescent simulator. By repeatedly generating samples of size n for K times, we obtained the summarized data points $D^k = \{(x^k(i), S^k(x(i))), 2 \leq x^k(i) \leq n, 1 \leq i \leq d\}, 1 \leq k \leq K$. Then, the NLS fitting algorithm described in previous section was applied to the parameter estimation for each D^k. The 95% confidence intervals of the parameters were estimated as the 2.5th and 97.5th quantiles from the K estimates.

Model Comparison

The well-known Akaike information criterion (AIC) is widely used for selection of models that better fit the observed data (Akaike 1974). AIC is based on the likelihood and the asymptotic properties of maximum-likelihood estimator, with the number of free parameters in the penalty term. A model with a smaller AIC is preferred among different models. As the full likelihood of our observed data is not known, we used a modified AIC in the spirit of the usual AIC. In our modified AIC, the likelihood is replaced by the quasi-likelihood which approximates the likelihood involving only first two moments of the data (McCullagh and Nelder 1989). Quasi-likelihood is often used for data without explicit distribution functions, and is often adopted in place of full likelihood for model selection (Pan 2001). Our modified AIC is defined as $\text{AIC} = -2\log L + 2p$, with p being the number of free parameters and $\log L$ the log quasi-likelihood. This modified AIC may not be the optimal criterion, but provides a ground for model comparison.

Let Γ denote parameters in mean and variance–covariance functions under a specific demographic model, and $\mathbf{S} = \{S_j(x_j), 1 \leq j \leq d\}$ denote the d data points from the TNSFS trajectory corresponding to subsample sizes $\mathbf{X} = \{x_j, 1 \leq j \leq d\}$. The quasi-likelihood then becomes

$$\log L(\Gamma \mid \mathbf{S}) = f(\mathbf{S} \mid \Gamma)$$
$$= \frac{1}{\sqrt{(2\pi)^d \mid \boldsymbol{\Sigma} \mid}} \exp\Big(-\frac{1}{2}(\mathbf{S} - \mu)^T \Sigma^{-1} (\mathbf{S} - \mu)\Big),$$

where Σ is the covariance matrix of S, $|\Sigma|$ is the determinant of Σ, each entry of $\boldsymbol{\mu}$, $\mu_j = f(x_j \mid \Gamma)$ is the expected TNS from a given sample x_j and the population growth model. The covariance matrix Σ takes into account the dependence among data points sampled from the TNSFS trajectories.

Note that in the above likelihood function, it is not trivial to obtain the exact form of the covariance matrix Σ. Therefore, we approximate Σ with the bootstrapping $\hat{\Sigma}$ estimated from data. Another consequence of using data-dependent covariance matrix is that in the quasi-likelihood of all the demographic models, the covariance matrices Σ are the same. We calculate AIC $= -2\log L + 2p$ as a criterion for model comparison in the "Application to Data" section.

Simulation Results

We examined the performance of our method using simulated data. A sample of 20,000 haplotypes was generated with the coalescent simulator ms for a 5-Mb region. We assumed the current haploid population size $N_0 = 2 \times 10^6$ and the population following a TEG with growth rates $r = 0.01, 0.05$, and 0.1 and onset time $\tau = 50$ and 100 generations. For each simulated data, a series of subsamples were randomly drawn from it, and the TNSFS trajectory was obtained. For each combination of the parameters, 40 samples were simulated, and the TNSFS trajectories were used as the input data for the NLS fitting method as described in previous section. We found that two parameters of the three, N_0, τ and r, can be estimated accurately and precisely if fixing the third. But if inferred jointly, they had some wider confidence intervals. Here we assumed the current population size N_0 was known, and investigated the performance of estimating the other two parameters. The boxplots of the inferred τ ad r were presented in figure 4. The method recovered the parameters accurately within the tested parameters ranges. We observed that when the growth rate was smaller, the estimates became biased, which required data of a larger sample size and spanning more regions to gain accuracy (results not shown).

The Effect of Sequencing Coverage on the Accuracy

The presented method assumes the TNSFS trajectories are generated from segregating sites without sequencing errors. In reality, sequencing data may be generated with low or mediate sequencing depth. The TNSFS trajectories are sensitive to the abundance of rare mutants, especially, the singletons and doubletons, which may be subject to high false positive rates in sequencing, and thus cause bias in estimate. To test the robustness of our method against sequencing error, we modified the above simulation pipeline to include the effect of sequencing error. For a chosen average sequencing depth, we assumed each position of the sequence (including both polymorphic sites and nonpolymorphic sites) had a number of sequencing reads, which follow a Poisson distribution with the mean equal to the sequencing depth C. For each of the sequencing reads, the sequencing error rate was chosen to be 0.5%. When the sequencing data were generated, we run samtools mpileup to determine the boundaries for the genotype calls for a given coverage level, and did genotype calling for each individual (Li et al. 2009). This SNP calling scheme performs worse than those Bayesian methods (such as, Li 2011; Nielsen et al. 2012), but is computationally inexpensive and works for large sample data in our case.

We simulated the demographic history with the parameters: $r = 0.05$, $\tau = 100$, and $N_0 = 2 \times 10^6$. We simulated four levels of sequencing depths: $C = 5, 10, 20,$ and 50. The inferred genotype data were used to construct the TNSFS trajectories, and the method was applied to the trajectories to infer the population growth parameters. The boxplots of the inferred parameters were presented in figure 5. When the sequencing depth is low ($C = 5$), the bias of inference is significant, but with the increase of sequencing depth (beyond $C = 10$), the bias becomes subtle (figs. 6 and 7).

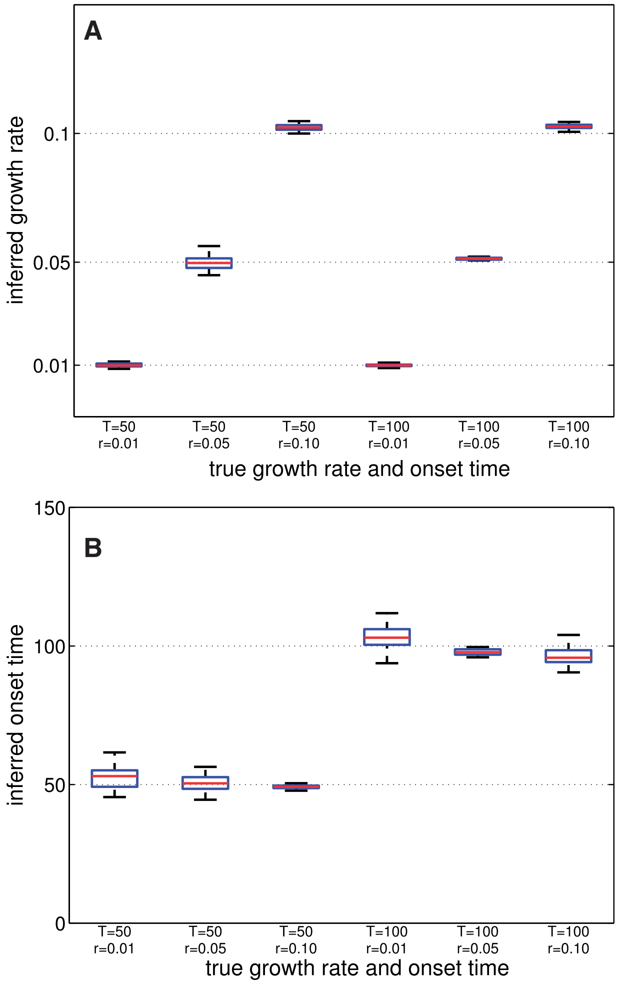

FIG. 4. The accuracy and precision of the estimated population growth rate and onset time of the exponential growth phase in the two-stage model. (A) Population growth rate. The X axis is the true growth rate used in coalescent simulation and the Y axis is the inferred growth rate values for 40 simulated samples. Each simulated sample was generated for a 5-Mb region with $n = 20, 000$, $N_0 = 2 \times 10^6$, and mutation rate $\mu = 10^{-8}$ per site per generation. (B) Population growth onset time. The other settings are the same as (A).

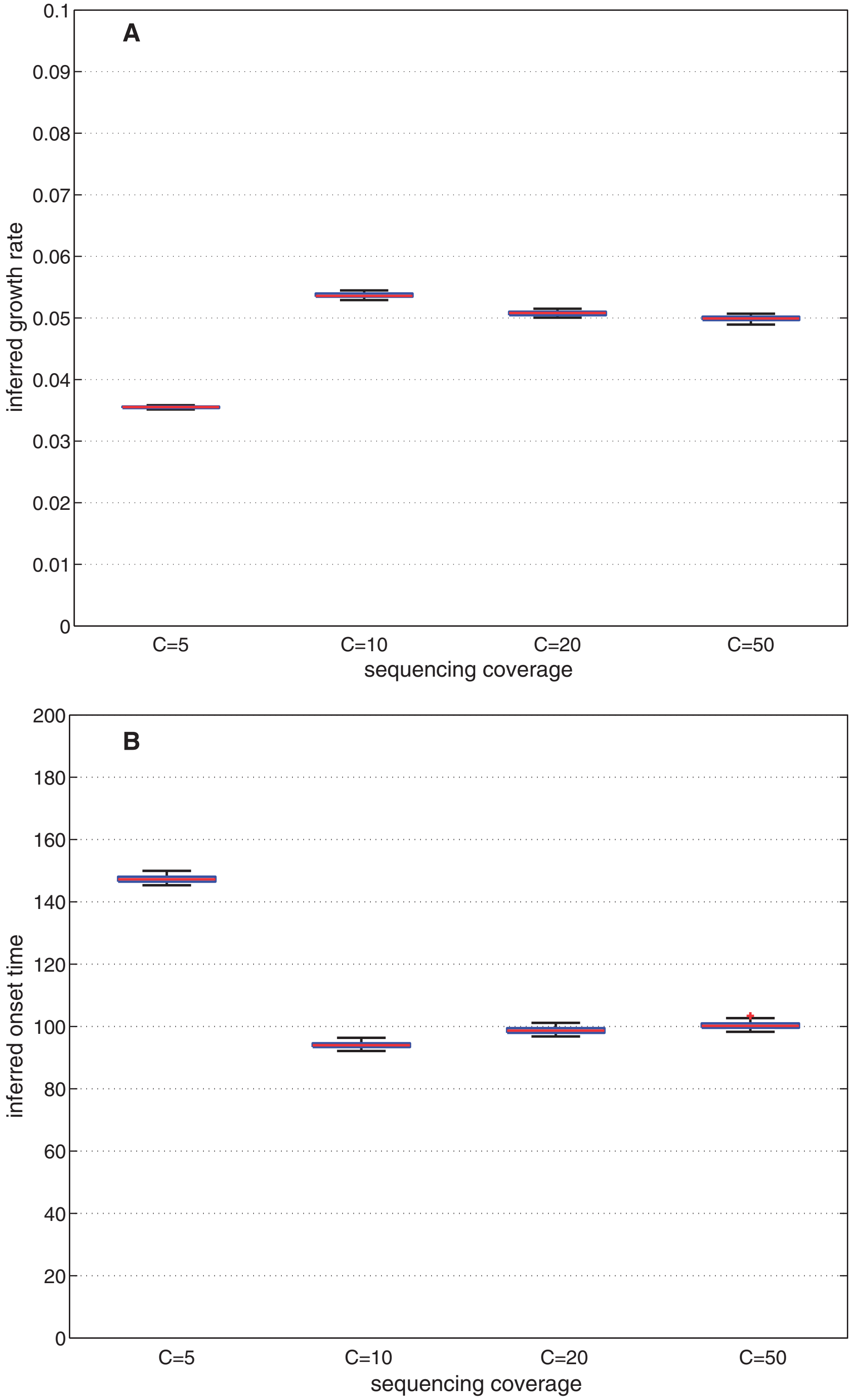

Fig. 5. The effect of sequencing coverage on the accuracy of parameter estimation of the exponential growth phase in the two-stage growth model. (*A*) Population growth rate. The *X* axis is the average sequencing coverage in the simulation and the *Y* axis is the inferred growth rate values for 40 simulated samples. Each simulated sample was generated for a 5-Mb region with $n = 20{,}000$, $N_0 = 2 \times 10^6$, mutation rate $\mu = 10^{-8}$ per site per generation, and the true growth rate is 0.05. (*B*) Population growth onset time. The true onset time is 100 generations ago. The other settings are the same as (*A*).

Application to Data

Exome Sequencing of European and African Populations

We applied the above proposed method to a recent population sequencing data set by Fu et al. (2013), in which 15,336 target genes for 4,298 individuals with European ancestry and 2, 217 individuals with African ancestry were sequenced. There are totally 709,816 and 643,128 SNPs identified in the two samples, respectively. We downloaded the vcf file from the NHLBI Exome Sequencing Project (ESP) Exome Variant Server (http://evs.gs.washington.edu/EVS/,

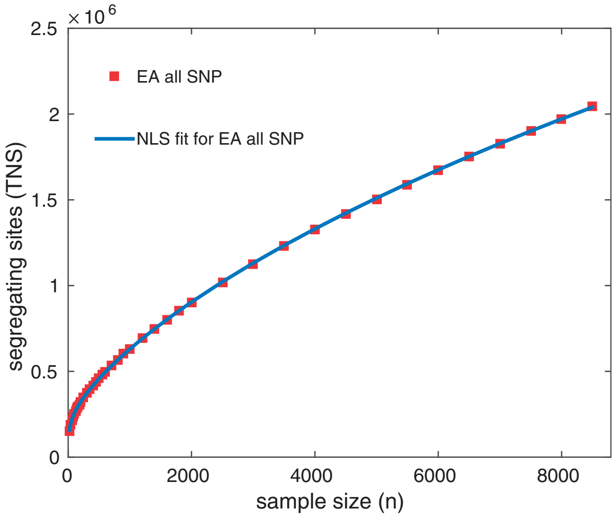

Fig. 6. The expected TNS ($\mathbb{E}S$) as a function of haploid sample size (n) for Europeans. The points represent data points from the ESP (Fu et al. 2013). The solid lines represent the fitted expectation of TNSFS with the population size N_0, the population growth rate r, and the initial time of the growth phase τ estimated by the nonlinear squares fitting.

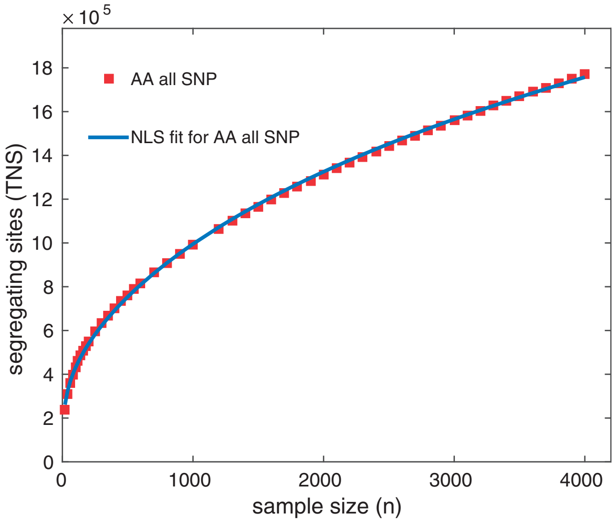

Fig. 7. The expected TNS ($\mathbb{E}S$) as a function of haploid sample size (n) for Africans. The points represent data points from the ESP (Fu et al. 2013). The solid lines represent the fitted expectation of TNSFS with the population size N_0, the population growth rate r, and the initial time of the growth phase τ estimated by the nonlinear squares fitting.

last accessed July 30, 2015). The TNSFS curves for a series of subsample sizes were constructed from the SNP allele frequencies in the vcf files.

We assumed a mutation rate $\mu = 1.2 \times 10^{-8}$ per site per generation (Genomes Project Consortium et al. 2010). We started with fitting a TEG model. We applied the NLS fitting method to the TNSFS of European populations, and estimated the starting time of the exponential growth of the European population to be 356 generations ago, or 8.9 ka assuming 25 years per generation, the growth rate $\hat{r} = 1.37\%$, and the current population size $\hat{N}_1 = 886,010$.

Table 2. The Comparison of Model Fitting on the European Population Sequencing Data of ESP.

Model	Inferred Parameters						−log*L*	*k*	AIC
	N_1	r	T_1	N_3	$N_4(N_6)$	N_5			
GM, 6 par	1,198,487	0.0149	290.4	2,020	13,143	62	44.59	6	101.18
GM, 5 par	1,131,347	0.0156	281.16	4,843	8,243	—	158.03	5	326.06
GM, 4 par	1,306,083	0.0176	225.0	—	10,332	—	5,497.56	4	11,003.0
GM, 3 par	1,224,400	0.0171	224.8	—	—	—	8,761.10	3	17,528.0
TEG, 3 par	886,010	0.0137	356.3	—	—	—	3,080.16	3	6,166.3

Note.—Figure 3 illustrates the parameters of the Gazave model. In the Gazave model, $N_4 = N_6$; $N_2 = N_1 e^{-rT_1}$. The nonfree parameters of the Gazave model in the table (indicated with "—") were fixed to the values from Gazave et al. (2014).

The TEG model may be oversimplified. We further explored more complicated models by taking into account the ancient demography. Recently, Gazave et al. (2014) proposed a six-epoch model for the European demographic history, which includes a recent exponential growth phase and five additional epochs with different population sizes and durations (fig. 3). The Gazave model is more realistic than the TEG model, as it explicitly models the bottlenecks during the Out-of-Africa migration and after the split of ancient Eurasian populations. Six parameters were included in the Gazave model: Three modeling the recent exponential growth phase, N_1, r, and T_1; and three additional, N_3, $N_4 = N_6$ and N_5, accounting for the population size jumps in the history. Here, N_3 is the population size of the recent population bottleneck, N_5 the population size during the ancient population bottleneck, N_4 and N_6 ancient population sizes that are set to be equal to reduce the number of parameters. Because the Gazave model is complicated and includes multiple parameters, Gazave et al. (2014) fixed some parameters at the values obtained from previous studies (Keinan et al. 2007), and only allowed three free parameters accounting for the recent exponential growth: N_1, r, and T_1 (or alternatively, N_1, T_1, and $N_2 = N_1 e^{-rT_1}$, see Model II in table 1 of Gazave et al. [2014]). In the following context, we referred to six-parameter Gazave model as 6p-GM, and three-parameter Gazave model as 3p-GM. We also modified 3p-GM to allow more free parameters.

We first fitted Model 6p-GM to the ESP data (fig. 6). The exponential growth stage of the European population was estimated to start at 290.4 (278–302) generations ago, or 7.26 ka (95% CI: 6.95 − 7.55ky) if assuming 25 years per generation, the growth rate estimated to be $\hat{r} = 1.49\%$ $(1.11 - 1.87\%)$, and the current population size $\hat{N}_1 = 1,198,487$ $(1.086 \times 10^6 - 1.312 \times 10^6)$. The AIC scores of the TEG model and the six-parameter GM indicate that Model 6p-GM performs better than TEG at data fitting ($\text{AIC}_{\text{TEG}} = 6,166.3$ vs. $\text{AIC}_{\text{6p-GM}} = 101.18$).

Next, we fixed some of the parameters for population size jumps at literature values. For example, in the five-parameter Gazave model (5p-GM), we set $N_5 = 189.4$ (the ancient bottleneck size) as in Gazave et al. (2014), and in the four-parameter Gazave model (4p-GM), we fixed one more parameter $N_3 = 549$ as in Gazave et al. (2014). The inferred parameter values and the corresponding AIC for Model 4p-GM and 5p-GM are listed in table 2. The comparison of AIC among various models we have fitted provides some interesting insights:

1) The substantial difference of AIC between 6p-GM and TEG demonstrates that ancient demography has a significant impact on estimating the parameters of recent population growth, which is consistent with the observation by Gazave et al. (2014).
2) The prominent improvement of model fitting from 4p-GM to 5p-GM by including N_3 indicates the significant effect of recent bottleneck on the TNSFS.
3) The improvement from 5p-GM to 6p-GM by including N_5, the ancient bottleneck is relatively minor. We also noticed that the three estimated parameters related to the recent population growth (N_1, r, and T_1) are quite similar between 5p-GM and 6p-GM.

We also inferred population growth for the African populations (fig. 7). African populations have a simpler demographic history, and we adopted the demographic model in Fu et al. (2013), which was modified based on Gravel et al. (2011). The model includes a recent exponential growth phase and two constant durations: The ancient population has a size of $N_3 = 7,310$, and it instantaneously became $N_2 = 14,474$ at 148 ka. We estimated the three free parameters of the exponential growth phase, and the ancient population size N_3 (this is also chosen based on model comparison using AIC scores, similarly to the analysis of the European data). The growth onset of the African populations was estimated to be $\hat{\tau} = 10.01$ ka (95% CI: $9.63 - 10.39$ky), with $\hat{r} = 0.735\%$ (95% CI: $0.593 - 0.877\%$), and the current effective population size $\hat{N_1} = 5.062 \times 10^5$ (95% CI: 4.293×10^5–5.831×10^5). The growth of the African populations is relatively mild compared with the European populations, and its onset time is consistent with the time of Agriculture origin around $8 - 10$ ka. Overall, our estimates of growth rates and onset times confirm former studies (Nelson et al. 2012; Tennessen et al. 2012).

Discussion

We derived the exact and asymptotic expectations of large-sample TNS for multiple-stage exponential growth models. The asymptotic equations are in simple analytical form and provide accurate approximation to the exact equations. Based on these asymptotic equations, an analog of Watterson's diversity measure θ_W can be derived for samples in nonequilibrium populations, or populations still undergoing expansion or contraction. Watterson's diversity measure ignores the influence of the demography on the genetic polymorphism, and thus leads to a bias estimate of the diversity for growing and contracting population. Our proposed diversity measure provides a consistent estimate of the genetic diversity.

We further demonstrated by simulations that the expected TBL derived under Kingman's coalescent framework is still valid for population level sequencing data that may violate Kingman's coalescent assumption $n << N$. Therefore, the expectation of TNS for large $n \to N$ can be approximated by that derived under Kingman's assumption.

Instead of using the full data or the AFS to construct the coalescent likelihood (Polanski and Kimmel 2003; Marth et al. 2004; Chen 2012), we focused on an informative statistic of the data: The increasing trend of the TNSFS. As the theoretical TNSFS is a nonlinear function of demographic parameters, such as the rate and the onset time of population growth, an NLS fitting to the observed TNSFS trajectory was used to infer population growth rate and onset time. We applied this method to simulated data and two real data sets to demonstrate its performance.

Our TNSFS method shares some similarities with the widely used NLFT methods (the number of lineages as a function of time, e.g., the skyline plot methods [Pybus et al. 2000; Drummond et al. 2005], and Maruvka et al. [2011]) in that both investigate the functional curves of summary statistics constructed from the sample. Our method has some good features in application compared with the NLFT methods: The functional curve of TNSFS analyzed by our method can be easily estimated from the sample, whereas the NLFT methods rely on the inference of a gene genealogy for the sample, which was not directly observable from the data and may introduce uncertainty if assuming a fixed genealogy is the true genealogy. Furthermore, the NLFT methods are more suitable for nonrecombining regions, such as, mitochondrial DNA and Y chromosomes, as it is reasonable to construct one single genealogy for one such region, but one single gene genealogy can hardly represent the histories across the recombining regions. Our method does not need construction of genealogies, and thus can be used to analyze data sampled from large-scale genomic regions with or without recombination.

It is known in population genetic studies that increasing the sample size has limited effect on the deep branches in a gene genealogy, and adding more individuals (haplotypes) to a sample mainly changes the lower part of a coalescent tree, implying that large samples have limited incremental benefits on inferring ancient demographic history comparing with small samples. On the other hand, as illustrated in this article and Coventry et al. (2010), a large sample of sequencing data helps to elucidate very recent events, such as the recent rapid growth of human populations, which was out of the scope of existing ancestral inference methods based on small samples (e.g., the PSMC method by Li and Durbin [2011]). Inferring very recent demographic history is important for medical studies, as demographic processes, as well as selection, shape the genetic architecture underlying diseases and complex traits. As the effective parameter range of the TNSFS method is different from the existing methods for ancient history, in the future study, we will combine the TNSFS method with other methods, such as the AFS-based methods, to jointly infer both ancient and very recent demography (Polanski and Kimmel 2003; Evans et al. 2007; Gutenkunst et al. 2009; Lukić et al. 2011; Živković and Stephan 2011; Chen 2012; Song and Steinrücken 2012; Harris and Nielsen 2013). We expect the two classes of methods will complement each other, and the method will be a very useful tool for population genetic inference based on large-scale population sequencing data.

Acknowledgments

The authors are grateful to Dr Joshua Akey, Dr Anand Bhaskar, Dr Heng Li, and Dr John Novembre for guidance in data analysis and simulations, to Dr Bing Su for insightful comments and hosting the visit to KIZ, to the editor and the anonymous reviewers for their valuable comments. This research was supported in part by the Hundred Talents Program of the Chinese Academy of Sciences (H.C.), the National Science Foundation through major research instrumentation grant number CNS-09-58854, and an NIH grant R01-GM078204 (J.H.).

Appendix A: The Expectation of Exact Intercoalescence Times for a Simple Exponential Growth Model

The exact distribution and expectation of intercoalescence times in a population with time-varying size were derived in Polanski et al. (2003) using integral transform. Here, we derive the expected coalescence times with a simple approach. Assume that n haplotypes are sampled from a contemporary population, and the population evolves deterministically with size $N(t)$ at t generations ago from the present. When tracing back in time, the exact probability that there are $A_n(t) = m$ ancestral haplotypes (lineages) at time t is (Griffiths and Tavaré 1998):

$$\begin{aligned} f(A_n(t) = m) &= p_{n,m}(t) \\ &= \sum_{i=m}^{n} A_i^{n,m} e^{\frac{-i(i-1)}{2N_0}\int_0^t \nu(u)\mathrm{d}u}, 1 \le m \le n, \end{aligned} \tag{A1}$$

where $N_0 \equiv N(0)$ is the current population size, and $\nu(t) = N_0/N(t)$. For a population under exponential growth with the rate r, $\nu(t) = e^{rt}$. The coefficient of the hypergeometric series is given by

$$A_i^{n,m} = \frac{(-1)^{i-m}(2i-1)m_{(i-1)}n_{[i]}}{m!(i-m)!n_{(i)}}, \tag{A2}$$

with $n_{(k)} = n(n+1)...(n+k-1)$ and $n_{[k]} = n(n-1)...(n-k+1)$ being the rising and falling factorial functions. The expectation of intercoalescence time W_m can then be derived using the distribution of lineage numbers (eq. A1) as:

$$\begin{aligned} \mathbb{E}W_m &= \int_{t=0}^{\infty} p_{n,m}(t)\mathrm{d}t = \int_{t=0}^{\infty} \sum_{i=m}^{n} A_i^{n,m} e^{-\frac{i(i-1)}{2N}\int_0^t e^{ru}\mathrm{d}u}\mathrm{d}t \\ &= \sum_{i=m}^{n} A_i^{n,m} e^{\frac{i(i-1)}{2Nr}} \int_0^{\infty} e^{-\frac{i(i-1)}{2Nr}e^{rt}}\mathrm{d}t \\ &= \sum_{i=m}^{n} A_i^{n,m} e^{\frac{i(i-1)}{2Nr}} \frac{1}{r}\mathrm{Ei}(-\frac{i(i-1)}{2Nr}), \end{aligned} \tag{A3}$$

where $\mathrm{Ei}(\cdot)$ denotes exponential integral (Press et al. 1992). It can be seen that equation (A3) is equivalent to equation (35) in Polanski et al. (2003).

Appendix B: The Total Branch Lengths of Gene Genealogies for a TEG Model

The TEG model was extensively used to approximate human demography in former studies (e.g., Adams and Hudson 2004; Evans et al. 2007; Coventry et al. 2010; Chen 2013). The model assumes that the ancient population has a constant size of N_a until time τ, and it starts an exponential growth since then with a rate of r until the present.

The derivation of the ETBLs of gene genealogies for such a two-stage growth model is more complicated than the simple exponential growth model presented in Appendix A. We divide the segregating sites S into two groups: Those arose in the ancient population before τ, S^a, and those arose during the growth phase, S^g. And similarly, the gene genealogy is also split into two parts, which are in the duration of the ancestral population, denoted by $\mathrm{TBL}^a(n)$, and in the growth phase, denoted by $\mathrm{TBL}^g(n)$. Here, (n) denotes that TBL is a function of the size of the contemporary sample. And $\mathbb{E}\mathrm{TBL}(n) = \mathbb{E}\mathrm{TBL}^g(n) + \mathbb{E}\mathrm{TBL}^a(n)$.

(1) $\mathrm{TBL}^g(n)$

If we assume the ancestral lineage number at time τ is m, then conditional on m, $\mathbb{E}(\mathrm{TBL}^g(n) \mid m)$ is

$$\mathbb{E}(\mathrm{TBL}^g(n) \mid m, \tau, r) = \sum_{j=m}^{n} j\mathbb{E}(W_j \mid m, \tau, r), \tag{A4}$$

where W_j is the intercoalescence time during which there are j lineages, r is the population growth rate, and τ is the onset time of the growth phase. Let $p_{n,m}(\tau)$ be the probability that there are m ancestral lineages at time τ, $\mathbb{E}\mathrm{TBL}^g(n)$ is obtained by summing over possible m, $1 \le m \le n$:

$$\mathbb{E}(\mathrm{TBL}^g(n)) = \sum_{m=1}^{n} p_{n,m}(\tau)\mathbb{E}(\mathrm{TBL}^g(n) \mid m, \tau, r). \tag{A5}$$

As in Chen (2013) the conditional expected intercoalescence time T_j given m, τ, r is written as:

$$\mathbb{E}(W_j \mid m, \tau, r) = \int_0^{\tau} \frac{p_{n,j}(t)p_{j,m}(\tau - t)}{p_{n,m}(\tau)}\mathrm{d}t. \tag{A6}$$

Substituting equations (A6) and (A4) into (A5), we get,

$$\begin{aligned} &\mathbb{E}(\mathrm{TBL}^g(n)) \\ &= \sum_{m=1}^{n} p_{n,m}(\tau) \sum_{j=m}^{n} j \int_0^{\tau} \frac{p_{n,j}(t)p_{j,m}(\tau - t)}{p_{n,m}(\tau)}\mathrm{d}t \\ &= \int_0^{\tau} \sum_{m=1}^{n}\sum_{j=m}^{n} jp_{n,j}(t)p_{j,m}(\tau - t)\mathrm{d}t \\ &= \int_0^{\tau} \sum_{j=1}^{n} jp_{n,j}(t) \sum_{m=1}^{j} p_{j,m}(\tau - t)\mathrm{d}t \end{aligned} \tag{A7}$$

$$
\begin{aligned}
&= \int_0^{\tau} \mathbb{E}A_n(t)\mathrm{d}t \\
&= \int_0^{\tau} \frac{n}{1+\frac{n}{2}g(t)}\mathrm{d}t \\
&= \frac{2nN_0(r\tau + \log(2N_0r) - \log((e^{r\tau}-1)n + 2N_0r))}{2N_0r - n}.
\end{aligned}
$$

(2) $\mathbf{TBL}^a(n)$

The part of a genealogy in the ancestral population before the onset of the growth phase is written as:

$$
\begin{aligned}
\mathbb{E}\mathrm{TBL}^a(n) &= \mathbb{E}_m(\mathbb{E}(\mathrm{TBL}^a(n)\,|\,m)) \\
&= \sum_{m=2}^{n} p_{n,m}(\tau)\mathbb{E}(\mathrm{TBL}^a(n)\,|\,m). \qquad (A8)
\end{aligned}
$$

$\mathbb{E}(\mathrm{TBL}^a(n)\,|\,m)$ is in well-known form if the ancient population is in equilibrium with a constant size before the exponential growth phase:

$$
\begin{aligned}
\mathbb{E}(\mathrm{TBL}^a(n)\,|\,m) &= 2N_a\sum_{j=1}^{m-1}\frac{1}{j} \qquad (A9)\\
&\approx 2N_a\log(m) + \gamma, \qquad (A10)
\end{aligned}
$$

with $\gamma = 0.57721566$ being the Euler constant. Equation (A10) is an accurate approximation of equation (A9) when the sample size m is large (Watterson 1975).

The expectation of $\mathrm{TBL}^a(n)$ is then estimated by

$$
\begin{aligned}
\mathbb{E}(\mathrm{TBL}^a(n)) &= \mathbb{E}(\mathbb{E}(\mathrm{TBL}^a(n)\,|\,m)) \\
&= 2N_a\mathbb{E}\left(\sum_{i=1}^{m-1}\frac{1}{i}\right) \\
&\approx 2N_a\mathbb{E}(\log(m) + \gamma) \qquad (A11)\\
&\approx 2N_a\left(\log(m) + \gamma - \frac{1}{2}(\frac{1}{u_m^2})\cdot\sigma_m^2\right) \\
&\approx 2N_a(\log(m) + \gamma).
\end{aligned}
$$

Note that in the above equation, u_m is the expected number of ancient lineages $A_n(\tau)$ at time τ, which is

$$
u_m = \mathbb{E}(A_n(\tau)) = \frac{n}{1 + n\frac{e^{r\tau}-1}{2Nr}}. \qquad (A12)
$$

The remaining term $\frac{1}{2}(\frac{1}{u_m^2})\cdot\sigma_m^2$ can be ignored since σ_m^2 is relatively very small when compared with u_m (Chen H and Chen K 2013), and the remaining term is on order m^{-1}, which shrinks to zero as $n \to \infty, n/m \to a$, and $a < \infty$.

Appendix C: The Total Branch Lengths of Gene Genealogies for a Three-Stage Exponential Growth Model

The three-stage exponential growth model can be written as:

$$
N(t) = \begin{cases} N_0e^{-r_1t},\ t \le \tau_1 \\ N_0e^{-r_1\tau_1 - r_2(t-\tau_1)},\ \tau_1 \le t \le \tau_2 \\ N_a,\ t > \tau_2. \end{cases} \qquad (A13)
$$

From this piecewise population size function, we can get the scaling function $g(t)$ for $t \le \tau_2$:

$$
g(t) = \begin{cases} \frac{1}{N_0r_1}(e^{r_1t} - 1),\ t \le \tau_1, \\ \frac{e^{r_1\tau_1}}{N_0r_2}(e^{r_2(t-\tau_1)} - 1) + \frac{1}{N_0r_1}(e^{r_1\tau_1} - 1),\ \tau_1 \le t \le \tau_2, \end{cases} \qquad (A14)
$$

and further we can obtain $\mathbb{E}\mathrm{TBL}^g$ for the two growth stages:

$$
\begin{aligned}
&\mathbb{E}\mathrm{TBL}^g \\
&= \frac{2nN_0}{2N_0r_1 - n}\left\{\begin{matrix} -\log(n(e^{r_1\tau_1} - 1) + 2N_0r_1) \\ +\log(2N_0r_1) + r_1\tau_1 \end{matrix}\right\} \\
&- \frac{2nN_0r_1}{ne^{r_1\tau_1}(r_2 - r_1) - nr_2 + 2N_0r_1r_2} \qquad (A15)\\
&\times \left\{\begin{matrix} \log\left(\begin{matrix} r_2(2N_0r_1 - n) + nr_1e^{r_1\tau_1 - r_2\tau_1 + r_2\tau_2} \\ +n(r_2 - r_1)e^{r_1\tau_1} \end{matrix}\right) \\ -\log(r_2(n(e^{r_1\tau_1} - 1) + 2N_0r_1)) \\ +r_2\tau_1 - r_2\tau_2 \end{matrix}\right\}.
\end{aligned}
$$

The $\mathbb{E}\mathrm{TBL}^a$ for the constant size stage is similar to that of the two-stage growth model shown in Appendix B.

Appendix D: The Total Branch Lengths of Gene Genealogies for the Gazave Model

Gazave et al. (2014) proposed a six-epoch model for the European population, which includes a phase of recent exponential growth, and five phases of constant sizes. Let M denote the number of epochs:

$$
N(t) = \begin{cases} N_1e^{-rt},\ t \le \tau_1, \\ N_k,\ \tau_{k-1} < t \le \tau_k,\ 2 \le k \le M. \end{cases} \qquad (A16)
$$

Then the time-scaling function is

$$
g(t) = \begin{cases} \frac{1}{N_1r}(e^{rt} - 1),\ t \le \tau_1, \\ \frac{1}{N_1r}(e^{r\tau_1} - 1) + \frac{t - \tau_1}{N_2},\ \tau_1 \le t \le \tau_2, \\ \frac{1}{N_1r}(e^{r\tau_1} - 1) + \sum_{j=2}^{k-1}\frac{\tau_j - \tau_{j-1}}{N_{k-1}} + \frac{t - \tau_{k-1}}{N_k}, \\ \qquad \tau_{k-1} \le t \le \tau_k,\ 2 < k \le M. \end{cases} \qquad (A17)
$$

The ETBL of the first phase (the exponential growth phase) follows (A7). The ETBL of the second phase is:

$$\mathbb{E}\mathrm{TBL}_2 = \int_{\tau_1}^{\tau_2} \mathbb{E}(A_n(t))\mathrm{d}t = \int_{\tau_1}^{\tau_2} \frac{n}{1+\frac{n}{2}g(t)}\mathrm{d}t$$

$$= 2N_2\{-\log((e^{r\tau}-1)/(N_1 r)+2/n)$$

$$+\log((e^{r\tau_1-1})/(N_1 r)+(\tau_2-\tau_1)/N_2+2/n)\} \qquad (A18)$$

Similarly, it can be easily shown that the ETBL of the k $(1 < k < M)$ stage of the Gazave model can be written as:

$$\mathbb{E}\mathrm{TBL}_k = 2N_k\log\left(\left\{\sum_{j=2}^{k}\frac{(\tau_j-\tau_{j-1})}{N_j}+\frac{(e^{r\tau_1}-1)}{rN_1}+\frac{2}{n}\right\}\right)$$

$$-2N_k\log\left(\sum_{j=2}^{k-1}\frac{\tau_j-\tau_{j-1}}{N_j}+\frac{(e^{r\tau_1}-1)}{rN_1}+\frac{2}{n}\right). \qquad (A19)$$

The ETBL of stage M follows equation (A11):

$$\mathbb{E}\mathrm{TBL}_M = 2N_M(\log(\mathbb{E}A_n(\tau_{M-1})+\gamma), \qquad (A20)$$

with $\mathbb{E}A_n(\tau_{M-1}) = \frac{n}{1+\frac{n}{2}g(\tau_{M-1})}$.

The total ETBL is the sum of the three components.

References

Adams A, Hudson R. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* 168(3):1699–1712.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr.* 19:716–723.

Altshuler D, Lander E, Ambrogio L, Bloom T, Cibulskis K, Fennell T, Gabriel S, Jaffe D, Shefler E, Sougnez C, et al. 2010. A map of human genome variation from population scale sequencing. *Nature* 467(7319):1061–1073.

Bhaskar A, Clark AG, Song YS. 2014. Distortion of genealogical properties when the sample is very large. *Proc Natl Acad Sci U S A.* 111(6):2385–2390.

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

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

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

Chen H, Green RE, Pääbo S, Slatkin M. 2007. The joint allele-frequency spectrum in closely related species. *Genetics* 177(1):387–398.

Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR, et al. 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun.* 1:131.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 22(5):1185–1192.

Efron B. 1982. The jackknife, the bootstrap, and other resampling plans. Vol. 38. Philadelphia: SIAM.

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

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson DA, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–220.

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

Fu YX. 2006. Exact coalescent for the Wright-Fisher model. *Theor Popul Biol.* 69(4):385–394.

Galassi M, Davies J, Theiler J, Gough B, Jungman G, Booth M, Rossi F. 2009. GNU Scientific Library Reference Manual. Bristol (United Kingdom): Network Theory.

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

Genomes Project Consortium, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.

Gravel S, Henn B, Gutenkunst R, Indap A, Marth G, Clark A, Yu F, Gibbs R, The 1000 Genomes Project, Bustamante C. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A.* 108(29):11983–11988.

Griffiths RC, Tavaré S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344:403–410.

Griffiths RC, Tavaré S. 1998. The age of a mutation in a general coalescent tree. *Stoch Models.* 14(1–2):273–295.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5(10):e1000695.

Harris K, Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.* 9(6):e1003521.

Hudson RR. 1990. Gene genealogies and the coalescent process. In: Futuyma DJ, Antonovics JD, editors. Oxford surveys in evolutionary biology. Vol. 7. New York: Oxford University Press. p. 1-44.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336(6082):740–743.

Keinan A, Mullikin JC, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet.* 39(10):1251–1255.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475(7357):493–496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and samtools. *Bioinformatics* 25(16):2078–2079.

Lukić S, Hey J. 2012. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* 192(2):619–639.

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

Marth GT, Czabarka E, Murvai J, Sherry ST. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 2004(166):351–372.

Maruvka Y, Shnerb N, Bar-Yam Y, Wakeley J. 2011. Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate. *Mol Biol Evol.* 28(5):1617–1631.

McCullagh P, Nelder JA. 1989. *Generalized linear models.* London: CRC Press.

Nei M, Chakraborty R, Fuerst PA. 1976. Infinite allele model with varying mutation rate. *Proc Natl Acad Sci U S A.* 73(11):4164–4168.

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

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. 2012. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One* 7(7):e37558.

Pan W. 2001. Akaike's information criterion in generalized estimating equations. *Biometrics* 57(1):120–125.

Polanski A, Bobrowski A, Kimmel M. 2003. A note on distributions of times to coalescence, under time-dependent population size. *Theor Popul Biol.* 63(1):33–40.

Polanski A, Kimmel M. 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* 165(1):427–436.

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

Pybus OG, Rambaut A, Harvey PH. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155(3):1429–1437.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.

Song YS, Steinrücken M. 2012. A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics* 190(3):1117–1129.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69.

Tyekucheva S, Makova KD, Karro JE, Hardison RC, Miller W, Chiaromonte F, et al. 2008. Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol.* 9(4):R76.

Wakeley J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J Mol Evol.* 37(6):613–623.

Wakeley J, Takahashi T. 2003. Gene genealogies when the sample size exceeds the effective size of the population. *Mol Biol Evol.* 20(2):208–213.

Watterson G. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7(2):256–276.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol.* 39(3):306–314.

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