Inference of Super-exponential Human Population Growth via Efficient Computation of the Site Frequency Spectrum for Generalized Models

Feng Gao[1] and Alon Keinan[1]
Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853

ABSTRACT The site frequency spectrum (SFS) and other genetic summary statistics are at the heart of many population genetic studies. Previous studies have shown that human populations have undergone a recent epoch of fast growth in effective population size. These studies assumed that growth is exponential, and the ensuing models leave an excess amount of extremely rare variants. This suggests that human populations might have experienced a recent growth with speed faster than exponential. Recent studies have introduced a generalized growth model where the growth speed can be faster or slower than exponential. However, only simulation approaches were available for obtaining summary statistics under such generalized models. In this study, we provide expressions to accurately and efficiently evaluate the SFS and other summary statistics under generalized models, which we further implement in a publicly available software. Investigating the power to infer deviation of growth from being exponential, we observed that adequate sample sizes facilitate accurate inference; *e.g.*, a sample of 3000 individuals with the amount of data expected from exome sequencing allows observing and accurately estimating growth with speed deviating by ≥10% from that of exponential. Applying our inference framework to data from the NHLBI Exome Sequencing Project, we found that a model with a generalized growth epoch fits the observed SFS significantly better than the equivalent model with exponential growth (P-value $= 3.85 \times 10^{-6}$). The estimated growth speed significantly deviates from exponential (P-value $\ll 10^{-12}$), with the best-fit estimate being of growth speed 12% faster than exponential.

KEYWORDS coalescent; generalized models; population growth; human demographic history; software

SUMMARY statistics of genetic variation play a vital role in population genetic studies, especially inference of demographic history. In particular, the site frequency spectrum (SFS) is a vital summary statistic of genetic data and is widely utilized by many demographic inference methods applied to humans and other organisms (Marth *et al.* 2004; Gutenkunst *et al.* 2009; Excoffier *et al.* 2013; Bhaskar *et al.* 2015; Liu and Fu 2015). Some other demographic inference methods are based on the sequential Markov coalescent and utilize the most recent common ancestor (T_{MRCA}) and linkage disequilibrium patterns (Li and Durbin 2011; Harris and Nielsen 2013; MacLeod *et al.* 2013; Sheehan *et al.* 2013; Schiffels and Durbin 2014). As another example, several studies used the average pairwise difference between chromosomes (Hammer *et al.* 2008; Gottipati *et al.* 2011; Arbiza *et al.* 2014) and the SFS (Keinan *et al.* 2009) to study the relative effective population sizes between the human X chromosome and the autosomes. The wide application of such genetic summary statistics stresses the need for their fast and accurate computation under any model of demographic history, instead of their estimations via simulations or approximations (*e.g.*, Hudson 2002; Gutenkunst *et al.* 2009).

Several recent demographic inference studies showed evidence that human populations have undergone a recent epoch of fast growth in effective population size (Gutenkunst *et al.* 2009; Coventry *et al.* 2010; Gravel *et al.* 2011; Nelson *et al.* 2012; Tennessen *et al.* 2012; Gazave *et al.* 2014). However, the above studies assumed that the growth is exponential. The observation of a huge amount of extremely rare,

doi: 10.1534/genetics.115.180570
Manuscript received July 7, 2015; accepted for publication September 28, 2015; published Early Online October 8, 2015.
Available freely online through the author-supported open access option.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.180570/-/DC1.
[1] Corresponding authors: Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853. E-mail: ak735@cornell.edu; and Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853. E-mail: fg237@cornell.edu

previously unknown variants in several sequencing studies with large sample sizes (Nelson *et al.* 2012; Tennessen *et al.* 2012; Fu *et al.* 2013) and the recent explosive growth in census population size suggests that the human population might have experienced a recent super-expononential growth, *i.e.*, growth with speed faster than exponential (Coventry *et al.* 2010; Keinan and Clark 2012; Reppell *et al.* 2012, 2014). Hence, recent studies presented a new generalized growth model that extends the previous exponential growth model by allowing the growth speed to be exponential or faster/slower than exponential (Reppell *et al.* 2012, 2014). Modeling the recent growth by this richer family of models holds the promise of a better fit to human genetic data and can also be applicable to other organisms that experienced growth. However, only simulation approaches are currently available for evaluating such a generalized growth demographic model (Reppell *et al.* 2012), which makes inference of demographic history computational intractable.

In this study, we first provide a set of explicit expressions for the computation of five summary statistics under a model of any number of epochs of generalized growth or decline: (1) the time to the most recent common ancestor (T_{MRCA}); (2) the total number of segregating sites (S); (3) the SFS; (4) the average pairwise difference between chromosomes per site (π); and (5) the burden of private mutations (α), a summary statistic that has been recently introduced as sensitive to recent growth (Keinan and Clark 2012; Gao and Keinan 2014). We also introduce a new software package, Efficient computation of Generalized models' Genetic summary Statistics (EGGS), which implements these expressions and facilitates fast and accurate generation of these summary statistics. We show that the numerically computed summary statistics match well with simulation results and facilitate computation that is orders of magnitude faster than simulations. By performing demographic inference on the SFS generated from simulated sequences, we then explore how many samples are needed for recovering parameters of a recent generalized growth epoch. Finally, we apply the software to investigate the nature of the recent growth in humans by inferring demographic models using the SFS of synonymous variants of 4300 European individuals from the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (Tennessen *et al.* 2012; Fu *et al.* 2013).

Materials and Methods

Generalized demographic models

A demographic model $N(T)$ describes the changes of effective population size N against time T. We consider time, measured in generations, as starting from 0 at present and increasing backward in time. Furthermore, we consider the families of demographic models that are constituted by any number of epochs of generalized growth or decline, along the lines of Bhaskar and Song (2014). More formally, there exists a minimal positive integer L such that the demographic history of a population can be split into a model with $L+1$ epochs that are split by L ordered different time points $T_1, T_2, \ldots, T_L$ ($T_0 = 0 < T_1 < T_2 < \ldots < T_L < T_{L+1} = \infty$), with the k^{th} epoch starting from T_{k-1} and lasting through T_k (thus the last epoch starts at time T_L and continues into indefinite past, $T_{L+1} = \infty$). Such a history is considered as a generalized model if the population size in each epoch $N(T_{k-1} \le T < T_k)$ can be described by the following differential equation regarding time T (Reppell *et al.* 2012, 2014),

$$\frac{dN}{dT} = -\, r_k N^{b_k}, \tag{1}$$

where $k = 1, 2, \ldots, L+1$. Each epoch can hence capture a variety of changing patterns in effective population size. Specifically, if $r_k = 0$, this epoch is of constant population size. When $r_k \neq 0$, b_k controls the growth or decline speed of this epoch: (1) if $b_k = 1$, the epoch is of exponential growth ($r_k > 0$) or decline ($r_k < 0$) with rate r_k; (2) if $b_k > 1$, the epoch is of faster-than-exponential (super-exponential) growth ($r_k > 0$) or decline ($r_k < 0$); (3) if $b_k < 1$, the epoch is of slower-than-exponential (sub-exponential) growth ($r_k > 0$) or decline ($r_k < 0$). Linear growth or decline is also a special case of generalized models when $b_k = 0$. An illustration of a generalized model with five epochs is provided in Figure 1, with more detailed explanation and illustrations in Supporting Information, File S1 and Figure S1.

The solution to Equation 1 is

$$N(T) = \begin{cases} \left(N_{k,\mathrm{i}}^{1-b_k} - r_k(T - T_{k-1})(1 - b_k)\right)^{\frac{1}{1-b_k}}, & b_k \neq 1 \\ N_{k,\mathrm{i}} e^{-r_k(T - T_{k-1})}, & b_k = 1 \end{cases} \tag{2}$$

(Reppell *et al.* 2012, 2014), where $N_{k,\mathrm{i}}$ is the initial population size of the k^{th} epoch. Each epoch k is defined by four parameters: the starting population size $N_{k,\mathrm{i}}$, the ending population size $N_{k,\mathrm{f}}$, the duration of the epoch $(T_k - T_{k-1})$, and the growth speed parameter b_k. The growth rate parameter r_k is an immediate function of these parameters, $r_k = r_k(N_{k,\mathrm{i}}, N_{k,\mathrm{f}}, b_k, T_k - T_{k-1})$, and hence does not need to be provided as an independent variable in defining the changes in effective population size during an epoch. Note that $N_{k+1,\mathrm{i}}$, the starting population size of the $(k+1)^{\text{th}}$ epoch, is not necessarily the same as $N_{k,\mathrm{f}}$, the ending population size of the k^{th} epoch. Specifically, if $N_{k+1,\mathrm{i}} \neq N_{k,\mathrm{f}}$, there is an instantaneous change in population size at time T_k.

Explicit expressions for summary statistics of demographic models under arbitrary population size functions

In this section, we briefly summarize the main results from previous studies that are used to evaluate the expected value of the summary statistics. Under Kingman's standard coalescent (Kingman 1982a,b), given a demographic model $N(T)$, the expected time to the most recent common ancestor $\mathbb{E}[T_{\mathrm{MRCA}}^{p}]$ can be calculated by

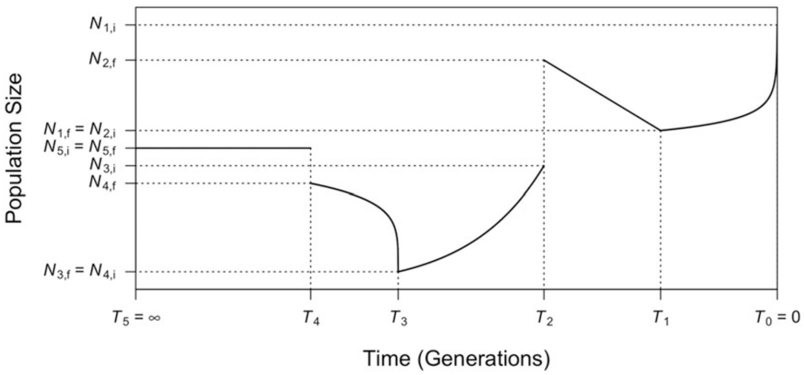

Figure 1 Illustration of an example of a generalized demographic model as introduced in the first section of *Materials and Methods*. This model consists of five epochs (starting from the present on the right): (1) faster-than-exponential ($b > 1$) growth (forward in time) from $N_{1,f}$ to $N_{1,i}$ between $T_0 = 0$ and T_1; (2) linear decline (a special case of generalized decline when $b = 0$) from $N_{2,f}$ to $N_{2,i}$ between T_1 and T_2; (3) exponential growth (a special case of generalized growth when $b = 1$) from $N_{3,f}$ to $N_{3,i}$ between T_2 and T_3; (4) slower-than-exponential ($b < 1$) decline from $N_{4,f}$ to $N_{4,i}$ between T_3 and T_4; and (5) constant population size (a special case of generalized growth when $r = 0$) at $N_{5,i} = N_{5,f}$ starting from T_4, which lasts indefinitely backward in time ($T_5 = \infty$). The ending population size of the previous epoch is not necessarily the beginning population size of the next epoch (*e.g.*, $N_{2,f} \neq N_{3,i}$, $N_{4,f} \neq N_{5,i}$), corresponding to an instantaneous population size change at that time.

$$\mathbb{E}\left[T^{p}_{\text{MRCA}}\right] = \sum_{j=2}^{p} A^{p}_{j}\psi_j \tag{3}$$

(Polanski and Kimmel 2003), where the superscript p is the number of chromosomes (*i.e.*, twice the sample size for diploids), ψ_j is the expected time to the first coalescent event when there are j chromosomes at present, and A^{p}_{j} are constants (Tavare 1984; Takahata and Nei 1985; Polanski *et al.* 2003) provided in File S1. Without loss of generality, we consider the case of diploid individuals, where there are $2N(T)$ chromosomes at any generation T, and use the notation $\mathcal{N}(T) = 2N(T)$. Then ψ_j is expressed by the equation

$$\begin{aligned}\psi_j &= \int_0^{\infty} T\frac{\binom{j}{2}}{\mathcal{N}(T)}e^{-\int_0^{T}\left(\binom{j}{2}d\sigma/\mathcal{N}(\sigma)\right)}dT \\ &= \int_0^{\infty} e^{-\binom{j}{2}\Lambda(T)}dT,\end{aligned} \tag{4}$$

where $\Lambda(T) = \int_0^{T}(d\sigma/\mathcal{N}(\sigma))$.

The expected full normalized SFS $\mathbb{E}[\boldsymbol{\xi}^{p}] = \left(\mathbb{E}[\xi^{p}_{1}], \mathbb{E}[\xi^{p}_{2}], \ldots, \mathbb{E}[\xi^{p}_{p-1}]\right)$ can be computed by the following set of equations (Polanski *et al.* 2003),

$$\mathbb{E}\left[\xi^{p}_{i}\right] = \frac{\mathbb{E}\left[\ell^{p}_{i}\right]}{\mathbb{E}[\mathcal{L}^{p}]};\quad \mathbb{E}\left[\ell^{p}_{i}\right] = \sum_{j=2}^{p} W^{p}_{i,j}\psi_j;\quad \mathbb{E}[\mathcal{L}^{p}] = \sum_{j=2}^{p} V^{p}_{j}\psi_j, \tag{5}$$

where ℓ^{p}_{i} is the length of branches in the genealogy that have i descendants ($i = 1, 2, \ldots, p-1$) and $\mathcal{L}^{p} = \sum_{i=1}^{p-1}\ell^{p}_{i}$ is the total length of all branches in the coalescent tree. The quantities V^{p}_{j} and $W^{p}_{i,j}$ are constants (Polanski *et al.* 2003), which we provide in File S1.

Naturally, the expected number of segregating sites is given by

$$\mathbb{E}[S] = \mu_0 L\mathbb{E}[\mathcal{L}^{p}], \tag{6}$$

where μ_0 is the mutation rate per site per generation and L is the length of the locus under consideration. The average pairwise difference between chromosomes per site $\mathbb{E}[\pi]$ can be calculated by

$$\mathbb{E}[\pi] = 2\mu_0\mathbb{E}\left[T^{p=2}_{\text{MRCA}}\right]. \tag{7}$$

The expected burden of private mutations α at a diploid sample size of $(p/2-1)$, defined as the proportion of heterozygous sites in a new diploid individual that are homozygous in the previous $(p/2-1)$ individuals, $\mathbb{E}[\alpha_{p/2-1}]$ can be computed by

$$\mathbb{E}\left[\alpha_{p/2-1}\right] = \frac{2}{p[1+\delta(1,p-1)]}\frac{\mathbb{E}\left[\ell^{p}_{1}\right] + \mathbb{E}\left[\ell^{p}_{p-1}\right]}{\mathbb{E}\left[\ell^{2}_{1}\right]} \tag{8}$$

(Gao and Keinan 2014), where $\delta(\cdot,\cdot)$ is Kronecker delta function.

The detailed description of the five summary statistics mentioned above is included in File S1.

Evaluation of the expected time to the first coalescent event under generalized models

The core of evaluating the summary statistics lies in finding feasible and numerically stable functions for calculating ψ_j, the expected time to the first coalescent event when there are j chromosomes at present. Previous studies give explicit expressions of ψ_j for a demographic model constructed by exponential and constant-size epochs (Polanski *et al.* 2003; Bhaskar *et al.* 2015). In this study, we give a comprehensive set of formulas for ψ_j under generalized models introduced above. Define $\phi^{k}_{j} := \int_{T_{k-1}}^{T_k} e^{-\binom{j}{2}\Lambda(T)}dT$; then $\psi_j = \sum_{k=1}^{L+1}\phi^{k}_{j}$, where $(L+1)$ is the total number of epochs. The quantity ϕ^{k}_{j} can be computed by the following set of equations:

1. If $r_k = 0$ or $b_k = 0, r_k \neq 0$,

$$\phi_j^k = \begin{cases} \dfrac{1}{\binom{j}{2}}\left[e^{-\binom{j}{2}\Lambda(T_k)}\mathcal{N}_{k,\mathrm{f}}\log\mathcal{N}_{k,\mathrm{f}} - e^{-\binom{j}{2}\Lambda(T_{k-1})}\mathcal{N}_{k,\mathrm{i}}\log\mathcal{N}_{k,\mathrm{i}} \right], \\ r_k + \binom{j}{2} = 0 \\ \dfrac{1}{r_k + \binom{j}{2}}\left[e^{-\binom{j}{2}\Lambda(T_{k-1})}\mathcal{N}_{k,\mathrm{i}} - e^{-\binom{j}{2}\Lambda(T_k)}\mathcal{N}_{k,\mathrm{f}} \right], \\ r_k + \binom{j}{2} \neq 0. \end{cases} \tag{9}$$

2. If $b_k > 0, r_k > 0$ or $b_k = 1, r_k < 0$,

$$\phi_j^k = \frac{1}{\binom{j}{2}}\left[\mathcal{N}_{k,\mathrm{i}}\mathcal{U}\left(2 - \frac{1}{b_k}, \frac{\binom{j}{2}}{b_k r_k}\mathcal{N}_{k,\mathrm{i}}^{-b_k}\right) e^{-\binom{j}{2}\Lambda(T_{k-1})} - \mathcal{N}_{k,\mathrm{f}}\mathcal{U}\left(2 - \frac{1}{b_k}, \frac{\binom{j}{2}}{b_k r_k}\mathcal{N}_{k,\mathrm{f}}^{-b_k}\right) e^{-\binom{j}{2}\Lambda(T_k)} \right]. \tag{10}$$

3. If $b_k < 0, r_k > 0$,

$$\phi_j^k = \frac{1}{\binom{j}{2}}\left[\mathcal{N}_{k,\mathrm{f}}\mathcal{M}\left(2 - \frac{1}{b_k}, \frac{\binom{j}{2}}{b_k r_k}\mathcal{N}_{k,\mathrm{f}}^{-b_k}\right) e^{-\binom{j}{2}\Lambda(T_k)} - \mathcal{N}_{k,\mathrm{i}}\mathcal{M}\left(2 - \frac{1}{b_k}, \frac{\binom{j}{2}}{b_k r_k}\mathcal{N}_{k,\mathrm{i}}^{-b_k}\right) e^{-\binom{j}{2}\Lambda(T_{k-1})} \right]. \tag{11}$$

The expressions of function $\Lambda(T)$ are given in File S1. The function $U(b,x) := xU(1,b,x) = x\int_0^\infty e^{-xt}(1+t)^{b-2}\,dt$, where $\mathcal{U}(a,b,x)$ is the confluent hypergeometric function of the second kind (Gradshteĭn *et al.* 2007). The function $M(b,x) := (x/(b-1))M(1,b,x) = x\int_0^1 e^{xt}(1-t)^{b-2}\,dt$, where $M(a,b,x)$ is the confluent hypergeometric function of the first kind (Gradshteĭn *et al.* 2007). The exponential growth or decline then becomes a special case of $\mathcal{U}(b,x)$ when $b = 1, x \neq 0$,

$$\mathcal{U}(1,x) = xe^x \int_1^\infty \frac{e^{-t}}{t}\,dt = xe^x E_1(x), \tag{12}$$

where $E_1(x)$ is the exponential integral (Gradshteĭn *et al.* 2007), which has been shown by previous studies (Polanski *et al.* 2003; Bhaskar *et al.* 2015). We could not find feasible and numerically stable closed-form formulas for ϕ_j^k when the population size decreases forward in time in a manner that is not linear or exponential (*i.e.*, $r_k < 0$ and $b_k \notin \{0, 1\}$). In these scenarios, we used Gauss–Legendre quadrature (Kahaner *et al.* 1988) for efficient numerical evaluation of relevant functions (see File S1 for detailed description).

Software implementation

The above expressions are implemented in a software package, EGGS. The source code and compiled programs for Linux and Mac OS platforms are publicly available from our Web site (http://keinanlab.cb.bscb.cornell.edu). Source code was written in C++, with no external libraries needed for compilation. Additional information of implementation is included in File S1 and in the manual that accompanies the software online.

Demographic models assumed in this study

The demographic models used in this study are based on the inferred European history presented by Gazave *et al.* (2014) (Figure 2, in black), which contains two bottlenecks (Keinan *et al.* 2007) and a recent exponential growth epoch. Specifically, the Gazave *et al.* (2014) model inferred that the European population had a constant effective population size of 10,000 (diploid) individuals before 4720 generations ago and went through the ancient bottleneck between 4720 and 4620 generations ago with a population size of 189. The population size then recovered to 10,000 diploids until 720 generations ago, at which time the recent bottleneck started with a size of 549. At 620 generations ago, the population size recovered to 5633 individuals. The recent growth epoch started 140.8 generations ago and led to a population size of 654,000 at present. The parameters of the original recent growth epoch were varied to incorporate generalized growth effects.

In addition to using the model mentioned above, we also applied an alternative model of ancient European history for inference. The model was first presented in Gravel *et al.* (2011) and later used in Tennessen *et al.* (2012). This model inferred that the European population had an ancient effective population size of 7300 diploid individuals until 6167 generations ago, when the population size expanded to 14,474 individuals. The first bottleneck took place 2125 generations ago, with the population size reducing to 1861 individuals. This first bottleneck lasted until 958 generations ago, at which time a second bottleneck took place with a

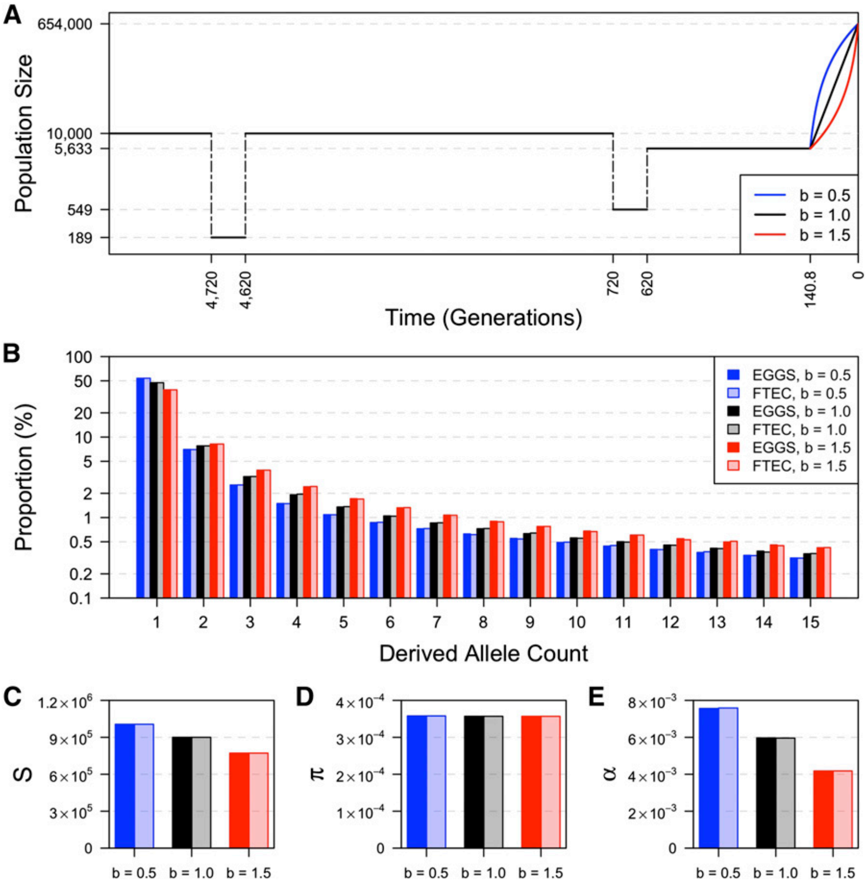

Figure 2 Comparison of four summary statistics estimated by FTEC simulation and computed by EGGS. (A) Demonstration of the demographic models considered for evaluating the accuracy of our calculations as implemented in EGGS (first section of *Results*). This two-bottleneck model has the same population size and time throughout history as in the inferred European history in Gazave *et al.* (2014), with the exception that we varied the growth speed parameter of the recent growth epoch to be $b = 0.5$ (sub-exponential, blue), $b = 1.0$ (exponential as in Gazave *et al.* 2014, black), and $b = 1.5$ (super-exponential, red). The *y*-axis shows effective population size of diploid individuals on log scale. (B–E) The comparison of the first 15 entries of the SFS (B), the total number of segregating sites (*S*) across all 200,000 loci (each 1000 bp long) (C), the expected pairwise difference between chromosomes per base pair (D), and the burden of private mutations (α) as the percentage of heterozygous variants in one individual that are monomorphic in the rest of the sample of 999 individuals (E) computed numerically in EGGS (dark-colored bars) and simulated by FTEC (light-colored bars) for the demographic models shown in A: blue, $b = 0.5$; black, $b = 1.0$; red, $b = 1.5$; with a sample size of 1000 individuals (2000 chromosomes). The *y*-axis in B is on a log scale.

decreased population size of 1032. We assumed 24 years per generation (Scally and Durbin 2012) to translate the year-based time presented in the original model. For compatibility with the Gazave *et al.* (2014) model, we considered that the population size had an instantaneous recovery after the second bottleneck lasted for 100 generations, instead of gradual recovery (Gazave *et al.* 2014). Figure S8 shows the schematic representation of the Gravel *et al.* (2011) model.

Demographic inference framework based on the site frequency spectrum

Demographic inference in this study was based on the observed allele frequency counts from the simulated or real data set. To determine the fitness of a model $N(T)$ to the observed data, we calculated the composite log likelihood by

$$\mathbb{L}[N] = \log \mathbb{E}[\boldsymbol{\xi}|N] = \mathbf{C} \cdot \mathbb{E}[\boldsymbol{\xi}|N], \qquad (13)$$

where $\mathbf{C}$ is a vector of the observed folded allele frequency counts and $\mathbb{E}[\boldsymbol{\xi}|N]$ is the computed folded SFS under demographic model $N(T)$. More detailed description can be found in File S1.

To search for the maximum-likelihood point over the parameter space, we applied the ECM (Expectation/Conditional Maximization) method (Meng and Rubin 1993), which was previously used in the demographic inference study by Excoffier *et al.* (2013). One hundred ECM cycles were performed for each run of inference. We obtained 95% confidence intervals of parameter estimates via block bootstrapping of the data 200 times. Specifically, if the original data contained l loci, we randomly chose l loci from the original data with replacement in each bootstrap (see File S1 for details).

Processing of NHLBI Exome Sequencing Project data for demographic history inference

The NHLBI Exome Sequencing Project (ESP) data (Tennessen *et al.* 2012; Fu *et al.* 2013) contain deep sequencing of 4300 individuals of European ancestry. An important feature of these data is the high level of sequencing coverage, which allows the capture of very rare variants accurately. These variants constitute the part of the SFS that is most enriched for information on recent population growth (Keinan and Clark 2012; Tennessen *et al.* 2012; Gao and Keinan 2014). To reduce the effect of selection as much as possible while keeping a sufficient amount of data, we chose to use the SFS calculated from synonymous single-nucleotide variants (SNVs) only, as previously performed by Tennessen *et al.* (2012). To further improve the quality of the data, we filtered SNVs with average read depth ≤20 or with successful genotype counts <7740 (90%) and subsampled the remaining 233,134 SNVs to 7740 alleles, which is equivalent to 3870 diploid individuals (File S1).

Data availability

The NHLBI Exome Sequencing Project (ESP) data used in this study is publicly available at http://evs.gs.washington.edu/EVS/.

Results

Comparison with simulated results by FTEC

To validate that the expressions provided in *Materials and Methods* can correctly compute the summary statistics under generalized growth models, we compared the summary statistics calculated by our software EGGS to those simulated by the software FTEC (a coalescent simulator for modeling faster than exponential growth by Reppell *et al.* 2012) under the demographic models shown in Figure 2A. This model is the inferred European history in Gazave *et al.* (2014), except that we varied the growth speed parameter b (Equation 1), which corresponds to 1 in the original model (exponential growth), to also be 0.5 (corresponding to sub-exponential growth) and 1.5 (corresponding to super-exponential growth). The sample size is fixed at 1000 diploid individuals (2000 chromosomes). For FTEC simulation, we used a mutation rate of 1.2×10^{-8} per base pair per generation (*e.g.*, Kong *et al.* 2012) and simulated 200,000 independent loci, each of 1000 bp.

The comparison of the SFS, S (across all 200,000 loci), π, and α numerically computed by EGGS to those simulated by FTEC is shown in Figure 2, B–E. For each demographic model illustrated in Figure 2A, the values for all summary statistics from the numerical computation by EGGS are practically identical to those from the simulation results by FTEC. However, our software EGGS exhibits a huge speed improvement over FTEC. For each model considered in Figure 2A, EGGS takes $<$1 sec to generate the results, while it takes $\sim$5 hr for FTEC to simulate the sequences, due to the large number of independent loci required for accurate estimation (performed in the Ubuntu system with an Intel Xeon CPU at 2.67 GHz). For instance, when 2000 independent loci are simulated, which still takes $\sim$3 min, the summary statistics deviate considerably from the accurate results (Figure S2 and Table S1). Furthermore, our software works well over a wide range of values of the growth parameter b, even when $b = 0$ (corresponding to linear growth or decline) or $b < 0$ (Figure S3), conditions that are not handled by FTEC. We note, however, that as a simulation program FTEC provides the full sequences as output and can have a wider range of applications than facilitated by the SFS and other summary statistics that EGGS calculates.

Evaluating inference of generalized growth based on the site frequency spectrum

We next set out to test the accuracy (as a function of sample size) of inferring parameters in models with generalized growth from the SFS. Bhaskar and Song (2014) showed that in theory, an underlying generalized growth demographic model can be uniquely identified by the ideal, perfect expected SFS with a very small sample size generated from that model (34 haploid sequences for the models shown in Figure 2A). However, the SFS is estimated in practice from a limited amount of data from each individual (even in the case of whole-genome sequencing) and, as a result, the estimated SFS will fluctuate around the expected values, which limits its accuracy for inference (Terhorst and Song 2015). We aim to test such inference in practice and determine the power of generalized growth detection and the sample size needed for accurately recovering the growth speed parameter as well as other parameters of the demographic model. For it to be comparable with many practical applications, we considered sequence length that is about equivalent to that obtained from whole-exome sequencing (File S1).

We performed inference on the SFS calculated from simulated sequences generated by FTEC. We simulated a demographic model with the same initial epochs as the model illustrated in Figure 2A. Starting 620 generations ago, the simulated model includes a constant population size of 10,000 until 200 generations ago, when the population starts a generalized growth epoch until the present. The generalized growth epoch starts with a population size of 10,000 that grows to an extant effective population size of 1 million individuals, with the growth speed parameter b taking each of the following values: 0.4, 0.7, 0.9, 1.0, 1.1, 1.3, and 1.6. We chose these values to represent a range of super-exponential and sub-exponential growth, with emphasis on values around the exponential rate ($b = 1.0$) to test the detection power of generalized growth when the growth speed deviates slightly from exponential. We varied the sample size (number of diploid individuals sampled at present) to be 1000, 2000, 3000, 5000, and 10,000 (File S1). The first 15 entries of the site frequency spectra for these simulated scenarios are shown in Figure S4. From each set of simulations, we then inferred four parameters of the recent growth epoch, which can uniquely determine the epoch: (1) the growth speed parameter b; (2) the initial population size before growth, N_{f}; (3) the ending population size after growth, N_{i}; and (4) the onset time of growth T, which is equivalent to the growth duration since the simulated epoch ends at present.

As sample size increases, the accuracy of the point estimates generally improves and the confidence interval narrows (Figure 3). Specifically, when the SFS of only 1000 diploids is used for inference, the inference performs poorly for all parameters, exhibiting large confidence intervals (Figure 3). However, the confidence interval always includes the true simulated value. A sample size of 2000 already exhibits acceptable performance except when the growth speed becomes large ($b = 1.3$ and 1.6). Larger sample sizes of 5000 and 10,000 are sufficient for inferring all parameters with very tight confidence intervals. For such sample sizes, the inference even significantly distinguishes between growth speeds ($b = 0.9$ and $b = 1.1$) that are close to exponential ($b = 1.0$) from that of an exponential, thereby concluding that a sub-exponential (0.9) or super-exponential

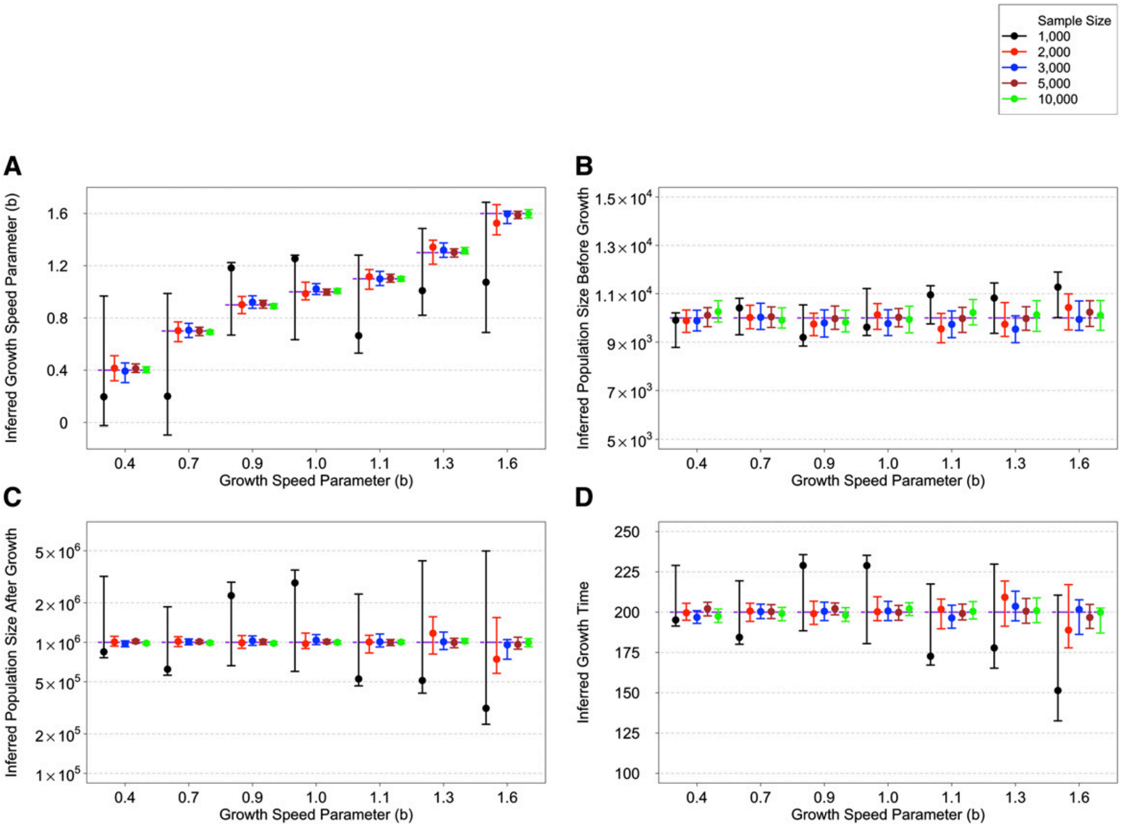

Figure 3 Inference results on simulated data with a recent generalized growth epoch. The model parameters are as follows: Growth starts 200 generations before the present from an effective population size of 10,000 and ends with an effective population size of 1 million at present. The growth speed parameter *b* takes the following values in different simulations: 0.4, 0.7, 0.9, 1.0, 1.1, 1.3, and 1.6. Inference of these four parameters is based on the SFS estimated from a sample of individuals of one of five sizes (black, 1000; red, 2000; blue, 3000; brown, 5000; and green, 10,000). The point estimates with 95% confidence interval for these parameters are grouped by the growth speed parameter *b* (*x*-axis). The thick, dashed lines show the true values of the simulated model. The results are shown in the following order: (A) the inferred growth speed parameter, (B) the inferred population size before growth, (C) the inferred population size after growth, and (D) the inferred growth start time. The *y*-axis in C is on a log scale.

(1.1) growth has taken place. These observations suggest that a sample size of at least 3000 diploid individuals might be needed for inferring the parameters associated with the simulated recent generalized growth epoch, which is motivated by previous models of European demographic history. It remains to be explored how accurate the estimates are, and how their accuracy improves with sample size, across a more diverse set of models.

European demographic history inference

We next performed demographic inference on NHLBI ESP data (Tennessen *et al.* 2012; Fu *et al.* 2013). We applied our inference framework to these data while considering and comparing two models. Both models assume the ancient epochs before 620 generations ago to be the same as those in the Gazave *et al.* (2014) model illustrated in Figure 2A. We inferred the parameters only for the most recent epoch, which is of *generalized* growth in one model while limited to *exponential* growth in the other. The parameters for inference are as follows: for both models, (1) population size before growth (N_f); (2) population size after growth (N_i); and (3) growth onset time (T), which is equivalent to the duration of growth; and only for the generalized growth model (4) the growth speed parameter (b), which is fixed at $b = 1$ for the exponential growth model. The point estimates and 95% confidence intervals are shown in Table 1 and the best-fit demographic models are illustrated in Figure 4, A and B (see also Figure S5, Figure S6 and Figure S7).

Although the Gazave *et al.* (2014) model assumed a different ancient history before the recent growth epoch from that assumed in Tennessen *et al.* (2012), using ESP data and assuming exponential growth, the inferred growth epoch is generally consistent with that obtained in the latter study (Figure 4, A and B, and Table 1). Our study infers that recent growth started 198 (95% C.I.: 195–202) generations ago with an effective population size of ~13,100 (12,600–13,600) and continued at a rate of 2.2% (2.15–2.26%) per generation (Table 1), while Tennessen *et al.* (2012) estimated that recent growth had an initial population size of ~9500 individuals, a duration of 204 generations, and a growth rate of 2.0% per generation.

The inferred generalized growth model fits the data significantly better than that with exponential growth (P-value $= 3.85 \times 10^{-6}$ by χ^2 likelihood-ratio test with 1 d.f.). It estimates that

Table 1 Demographic inference results using ESP data for a model with a recent epoch of exponential growth and a model with a recent epoch of generalized growth

Ancient history	Growth model	N_f (10^4)	N_i (10^6)	T	b
Gazave model	Exponential	1.31 (1.26–1.36)	1.04 (1.00–1.07)	198 (195–202)	NA
	Generalized	1.24 (1.18–1.30)	1.26 (1.16–1.37)	213 (206–220)	1.12 (1.07–1.15)
Gravel model	Exponential	0.89 (0.86–0.93)	0.85 (0.82–0.88)	186 (182–190)	NA
	Generalized	0.78 (0.74–0.83)	1.33 (1.22–1.46)	218 (211–228)	1.22 (1.18–1.26)

Shown are point estimates and 95% confident intervals (in parentheses) for the following parameters of the inferred recent growth epoch when the ancient history was assumed to be the same as that in the Gazave *et al.* (2014) model and the Gravel *et al.* (2011) model: population size before growth (N_f); population size after growth (N_i); time growth started in generations (T); and the growth speed parameter (b), which is fixed at $b = 1$ in the exponential growth case.

growth started 213 (206–220) generations ago from an effective population size of 12,400 (11,800–13,000), both values consistent with those estimated in the exponential growth model. The extant effective population size following growth is estimated to be 1.26 (1.16–1.37) million. The inferred growth speed parameter $b = 1.12$ (1.07–1.15) is significantly larger than the exponential speed of $b = 1$ (P-value $\ll 10^{-12}$, using a one-tailed z-test), which is the main difference between the two models. $b = 1.12$ implies a growth rate acceleration pattern (File S1) that is super-exponential at 12% faster than exponential through the epoch (Figure 4): the super-exponential growth is relatively slow around the onset time, and it keeps accelerating as time approaches the present.

To test the sensitivity of the model to the assumption of ancient European history, we considered an alternate model of ancient history. We fixed the history before 858 generation ago to be that inferred by Gravel *et al.* (2011) for Europeans (*Materials and Methods*). We repeated inference of the same parameters, using the same ESP data. As above, the inferred parameters for exponential growth are similar to those obtained in Tennessen *et al.* (2012) that were based on the model of Gravel *et al.* (2011) (Table 1). However, the SFS from this model fits the data worse than that from the exponential model based on the ancient history of the Gazave *et al.* (2014) model (P-value $= 1.59 \times 10^{-6}$ from χ^2 goodness-of-fit test between the exponential Gravel model and ESP data; P-value $= 0.97$ for the corresponding exponential Gazave model; see File S1 and Table S3). By applying a generalized growth epoch to the Gravel *et al.* (2011) model, the inferred parameters are generally in line with those from the generalized model based on Gazave *et al.* (2014), although some differences exist (Table 1), indicating that the assumption of ancient history can affect the inference of recent growth to some extent. More importantly, the generalized Gravel model fits the data almost equally well as the generalized Gazave model, which is significantly better than the exponential model (P-value $\ll 10^{-12}$ by χ^2 likelihood-ratio test; also see Table S3). As with the generalized Gazave model, the inferred growth speed parameter from the generalized Gravel model, $b = 1.22$ (1.18–1.26), is also significantly larger than the exponential speed $b = 1$ (P-value $\ll 10^{-12}$, using a one-tailed z-test; Figure 4, C and D).

Motivated by these results, we considered a third model with *two* recent exponential growth epochs, which still assumes the ancient epochs before 620 generations ago to be the same as those in the Gazave *et al.* (2014) model illustrated in Figure 2A. Five parameters were inferred (Table S2), with the first phase of growth estimated to start 219 (95–334) generations ago with a population size of 12,200 (11,700–13,200). This phase of growth lasts until 135 (25–157) generations ago and leads to a population size of 47,100 (30,200–540,900). The population size after the recent phase of growth is 1.12 (1.07–2.09) million. This model provides a significantly better fit than the model with a single exponential growth (P-value $= 5.55 \times 10^{-6}$ by χ^2 likelihood-ratio test with 2 d.f.), but is a worse model than the generalized growth model (based on the Bayesian information criterion, $\mathrm{BIC}_{\text{two-epoch exponential}} - \mathrm{BIC}_{\text{generalized}} = 6.1$). However, this model exhibits some of the same accelerating patterns as in the generalized growth model, ascertained by the growth rate of the most recent exponential epoch being 2.4% (2.3–5.2%), larger than that of the first exponential epoch, 1.6% (1.3–2.1%). This acceleration pattern shown in both the generalized model and the model with two exponential epochs is consistent with evidence of growth in European census population size that has greatly accelerated in the modern era (Keinan and Clark 2012).

Discussion

In this study, we provide mathematical derivation and a software that can efficiently compute the expected values of five genetic data summary statistics given a generalized demographic model by evaluating the derived explicit expressions. These summary statistics include the time to the most recent common ancestor (T_{MRCA}), the total number of segregating sites (S), the SFS, the average pairwise difference between chromosomes per site (π), and the burden of private mutations (α). The fast and accurate generation of these summary statistics under generalized models can provide a useful tool in the studies of human demographic inference. For instance, in addition to inference based on the SFS as in the present study, a recent study by Chen *et al.* (2015) presented an inference framework based on the total number of segregating sites. The results in this study can be easily incorporated into that framework. Furthermore, the source code of the software is freely available to allow extensions to compute other summary statistics of interest (for example, the joint SFS of samples from multiple populations under

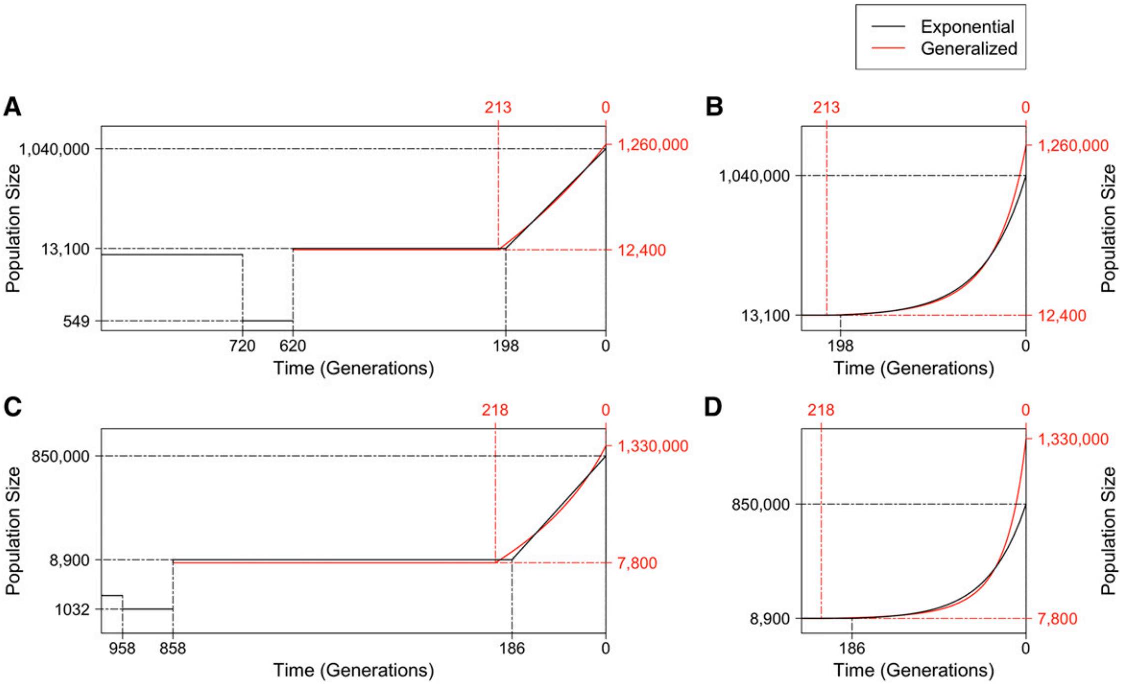

Figure 4 Demographic inference results based on ESP data. (A) Illustration of the effective population size (*y*-axis, on a log scale) over time for the best-fit models inferred based on ESP data, assuming the ancient history is the same as that in Gazave *et al.* (2014). Two models are shown: one restricted to recent growth being exponential (black) and one with a generalized recent growth epoch (red). Before 620 generations ago, the model was not inferred and all parameters were set to be the same as those shown in Figure 2A. Solid lines show the effective population size over time for each of the inferred models, with dashed lines indicating estimated parameter values on the *x*-axis or the *y*-axis. Only the most recent 1000 generations are shown to emphasize the difference between the two models. (B) A zoom-in to the most recent 240 generations of the inferred models in A to emphasize the acceleration pattern of the generalized growth model, with the *y*-axis on a linear scale. (C-D) Similar to A-B, except that the best-fit models presented are based on the assumption that the ancient history before 858 generations ago is fixed to that in Gravel *et al.* (2011) (see Figure S8).

generalized models, by extending the work of Wakeley and Hey 1997 and Chen 2012). Such extensions can facilitate a variety of population genetic studies in humans and other organisms beyond the inference of demographic history.

It is also possible that other families of growth models may fit the pattern of human population size history. For instance, Eldon *et al.* (2015) considered the algebraic-growth model in the form of $N(T) = T^{\gamma}$. In reality, however, not all demographic models have numerically stable closed-form expressions for the expected time to the first coalescent event (ψ_j). In these cases, fast and accurate numerical integration methods, such as the Gauss–Legendre quadrature used in this work, can be applied to evaluate ψ_j. This technique holds the promise of efficiently generating the expected value of population genetic summary statistics under arbitrary population size functions.

Bhaskar *et al.* (2014) pointed out that as sample size increases, the assumptions of standard Kingman's coalescent are violated as multi-merger and simultaneous-merger events can become nonnegligible. Such events can distort the genealogies and potentially cause the values of summary statistics to be different from those under Kingman's coalescent (Bhaskar *et al.* 2014). To explore such discrepancies, we compared the SFS from Kingman's coalescent and the discrete-time Wright–Fisher (DTWF) model (Bhaskar *et al.* 2014) under the inferred demographic history in the generalized Gazave model with a sample size of 3870 diploids (File S1). We observed that the SFS from the DTWF model and Kingman's coalescent are very similar (File S1 and Figure S9), which means that multi-merger and simultaneous-merger events should not have a significant effect on the inference carried out in this study. However, it remains valuable to systematically study the effect of multi-merger and simultaneous-merger events in the context of generalized growth, especially as sample size increases.

By applying inference of generalized growth based on the SFS generated from the synonymous variants of 4300 individuals of the NHLBI ESP data set (Tennessen *et al.* 2012; Fu *et al.* 2013), we found that the generalized growth model shows a better fit to the observed data than the exponential growth model that has been used by almost all previous demographic modeling studies (P-value $= 3.85 \times 10^{-6}$). We also found that the European population experienced a recent growth in population size with speed modestly faster than exponential ($b = 1.12$, P-value $\ll 10^{-12}$ for difference from $b = 1$). This result is consistent with previous speculations that the human population might have undergone a recent accelerated growth epoch based on the observation of very rare, previously unknown variants in several sequencing studies with large sample sizes (Nelson *et al.* 2012;

Tennessen *et al.* 2012; Fu *et al.* 2013). It is also in line with the super-exponential growth in census population size during that time (Keinan and Clark 2012). In future studies, it will be valuable to incorporate gradient-based optimization techniques for the fast inference of demographic models containing generalized growth epochs, *e.g.*, by extending the work of Bhaskar *et al.* (2015). Such an improvement will enable simultaneous inference of recent growth and more ancient epochs.

To minimize the impact of natural selection on our demographic inference, we considered only synonymous SNVs for demographic modeling, as in the original study of Tennessen *et al.* (2012). However, it is still a potential limitation that the data are affected by negative and background selection. Hence, it remains valuable to validate the result of super-exponential growth by conducting inference on SFS calculated from more neutral genomic regions (Gazave *et al.* 2014) or by modeling the effect of selection. One promising possibility is extracting genomic regions that are less subject to selection from whole-genome sequences in the UK10K project (The UK10K Consortium *et al.* 2015). More generally, with the increasing availability of high-quality whole-genome sequencing data with large sample sizes for humans and other species, more refined and realistic demographic histories can be estimated with generalized models.

Acknowledgments

The authors thank Leonardo Arbiza for helpful comments; Yun S. Song, Andrew G. Clark, and two anonymous reviewers for insightful comments on earlier versions of this manuscript; and Arjun Biddanda for his careful editing of the software manual. This work was supported by National Institutes of Health grants R01GM108805 and R01HG006849, an award from The Ellison Medical Foundation, and an award from The Edward Mallinckrodt, Jr. Foundation. Feng Gao is a Howard Hughes Medical Institute International Student Research fellow.

Literature Cited

Arbiza, L., S. Gottipati, A. Siepel, and A. Keinan, 2014 Contrasting X-linked and autosomal diversity across 14 human populations. Am. J. Hum. Genet. 94(6): 827–844.

Bhaskar, A., and Y. S. Song, 2014 Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. Ann. Stat. 42(6): 2469–2493.

Bhaskar, A., A. G. Clark, and Y. S. Song, 2014 Distortion of genealogical properties when the sample is very large. Proc. Natl. Acad. Sci. USA 111(6): 2385–2390.

Bhaskar, A., Y. X. Wang, and Y. S. Song, 2015 Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Res. 25(2): 268–279.

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

Chen, H., J. Hey, and K. Chen, 2015 Inferring very recent population growth rate from population-scale sequencing data: using a large-sample coalescent estimator. Mol. Biol. Evol. 32(11): 2996–3011.

Coventry, A., L. M. Bull-Otterson, 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.

Eldon, B., M. Birkner, J. Blath, and F. Freund, 2015 Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? Genetics 199: 841–856.

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

Fu, W., T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis *et al.*, 2013 Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493(7431): 216–220.

Gao, F., and A. Keinan, 2014 High burden of private mutations due to explosive human population growth and purifying selection. BMC Genomics 15(Suppl. 4): S3.

Gazave, E., L. Ma, D. Chang, A. Coventry, F. Gao *et al.*, 2014 Neutral genomic regions refine models of recent rapid human population growth. Proc. Natl. Acad. Sci. USA 111(2): 757–762.

Gottipati, S., L. Arbiza, A. Siepel, A. G. Clark, and A. Keinan, 2011 Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. Nat. Genet. 43(8): 741–743.

Gradshteĭn, I. S., I. M. Ryzhik, and A. Jeffrey, 2007 *Table of Integrals, Series, and Products*, Ed. 7. Academic Press, Amsterdam/Boston.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth *et al.*, 2011 Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA 108(29): 11983–11988.

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(10): e1000695.

Hammer, M. F., F. L. Mendez, M. P. Cox, A. E. Woerner, and J. D. Wall, 2008 Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genet. 4(9): e1000202.

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

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18(2): 337–338.

Kahaner, D., C. B. Moler, S. Nash, and G. E. Forsythe, 1988 *Numerical Methods and Software*. Prentice Hall, Englewood Cliffs, NJ.

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

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

Keinan, A., J. C. Mullikin, N. Patterson, and D. Reich, 2009 Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat. Genet. 41(1): 66–70.

Kingman, K. F. C., 1982a On the genealogy of large populations. J. Appl. Probab. 19: 27–43.

Kingman, K. F. C., 1982b The coalescent. Stoch. Proc. Appl. 13(3): 235–248.

Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem *et al.*, 2012 Rate of de novo mutations and the importance of father's age to disease risk. Nature 488(7412): 471–475.

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

Liu, X., and Y. X. Fu, 2015 Exploring population size changes using SNP frequency spectra. Nat. Genet. 47(5): 555–559.

MacLeod, I. M., D. M. Larkin, H. A. Lewin, B. J. Hayes, and M. E. Goddard, 2013 Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. Mol. Biol. Evol. 30(9): 2209–2223.

Marth, G. T., E. Czabarka, 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 166: 351–372.

Meng, X. L., and D. B. Rubin, 1993 Maximum-likelihood-estimation via the Ecm algorithm - a general framework. Biometrika 80(2): 267–278.

Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337(6090): 100–104.

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(1): 33–40.

Reppell, M., M. Boehnke, and S. Zollner, 2012 FTEC: a coalescent simulator for modeling faster than exponential growth. Bioinformatics 28(9): 1282–1283.

Reppell, M., M. Boehnke, and S. Zollner, 2014 The impact of accelerating faster than exponential population growth on genetic variation. Genetics 196: 819–828.

Scally, A., and R. Durbin, 2012 Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. 13(10): 745–753.

Schiffels, S., and R. Durbin, 2014 Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46(8): 919–925.

Sheehan, S., K. Harris, and Y. S. Song, 2013 Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics 194: 647–662.

Takahata, N., and M. Nei, 1985 Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110: 325–344.

Tavare, S., 1984 Line-of-descent and genealogical processes, and their applications in population-genetics models. Theor. Popul. Biol. 26(2): 119–164.

Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090): 64–69.

The UK10K Consortium, K. Walter, J. L. Min, J. Huang, L. Crooks *et al.*, 2015 The UK10K project identifies rare variants in health and disease. Nature 526(7571): 82–90.

Terhorst, J., and Y. S. Song, 2015 Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. Proc. Natl. Acad. Sci. USA 112(25): 7677–7682.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics 145: 847–855.

Communicating editor: S. Ramachandran

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.180570/-/DC1

Inference of Super-exponential Human Population Growth via Efficient Computation of the Site Frequency Spectrum for Generalized Models

Feng Gao and Alon Keinan

DOI: 10.1534/genetics.115.180570

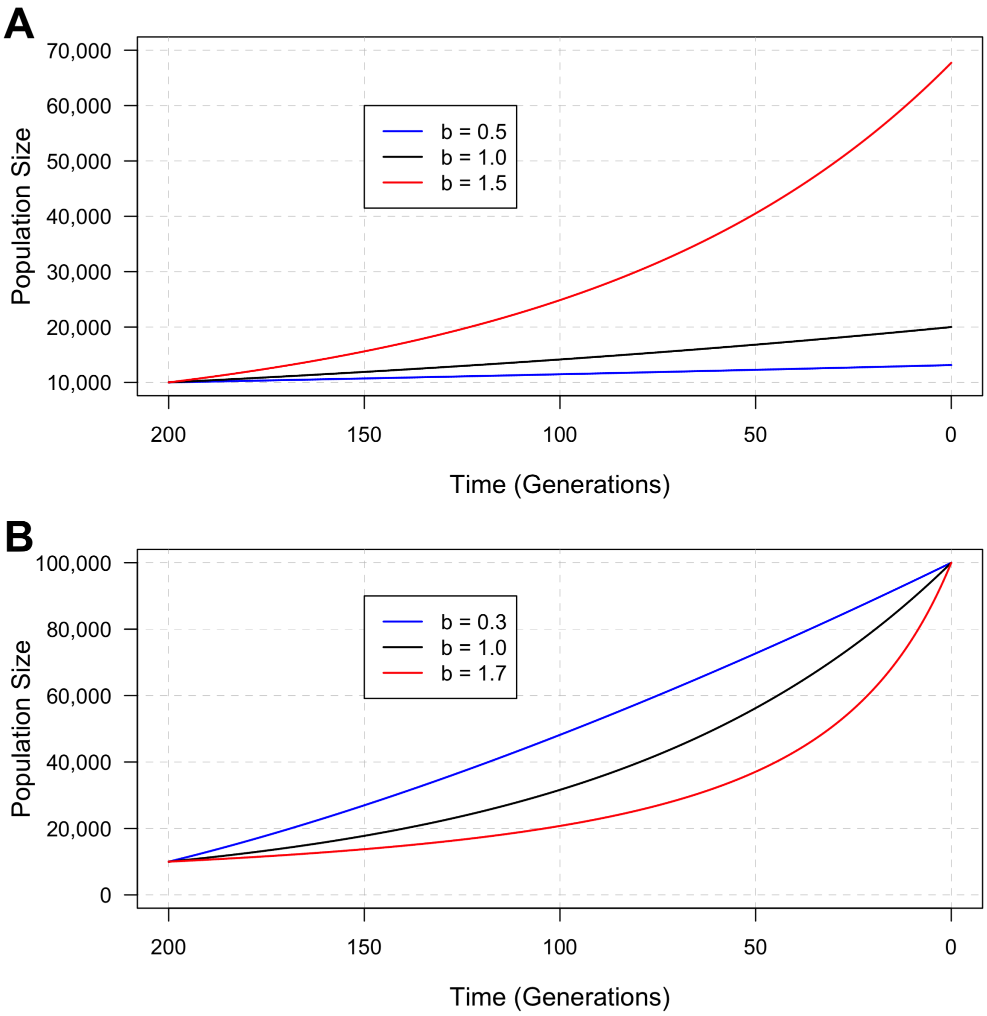

Figure S1. Different patterns of generalized growth. (A) Illustration of the population size functions when keeping the population size before growth N_f, the growth time T and the parameter r the same and varying the growth speed parameter b to be 0.5, 1.0 and 1.5. (B) Illustration of the population size functions when keeping the population size before growth N_f, the population size after growth N_i and the growth time T the same and varying the growth speed parameter b to be 0.3, 1.0 and 1.7.

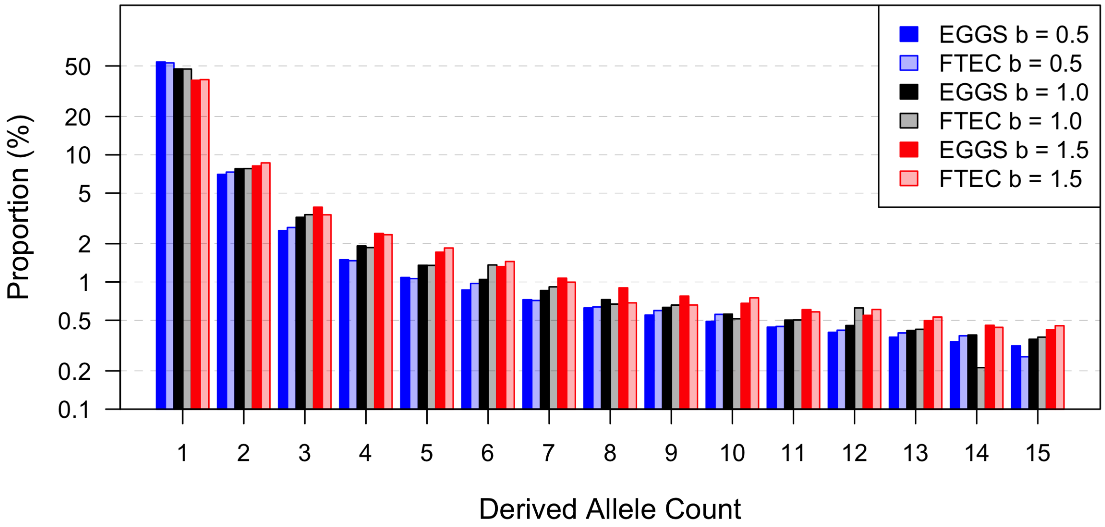

Figure S2. Comparison of the first 15 entries of the SFS computed numerically in EGGS **(dark bars) and simulated result by** FTEC **(light bars).** Only 2,000 loci (1,000 bp-long each) instead of 200,000 were simulated for the demographic models shown in Figure 2(A): $b = 0.5$, blue; $b = 1.0$, black; $b = 1.5$, red. y-axis is on log scale.

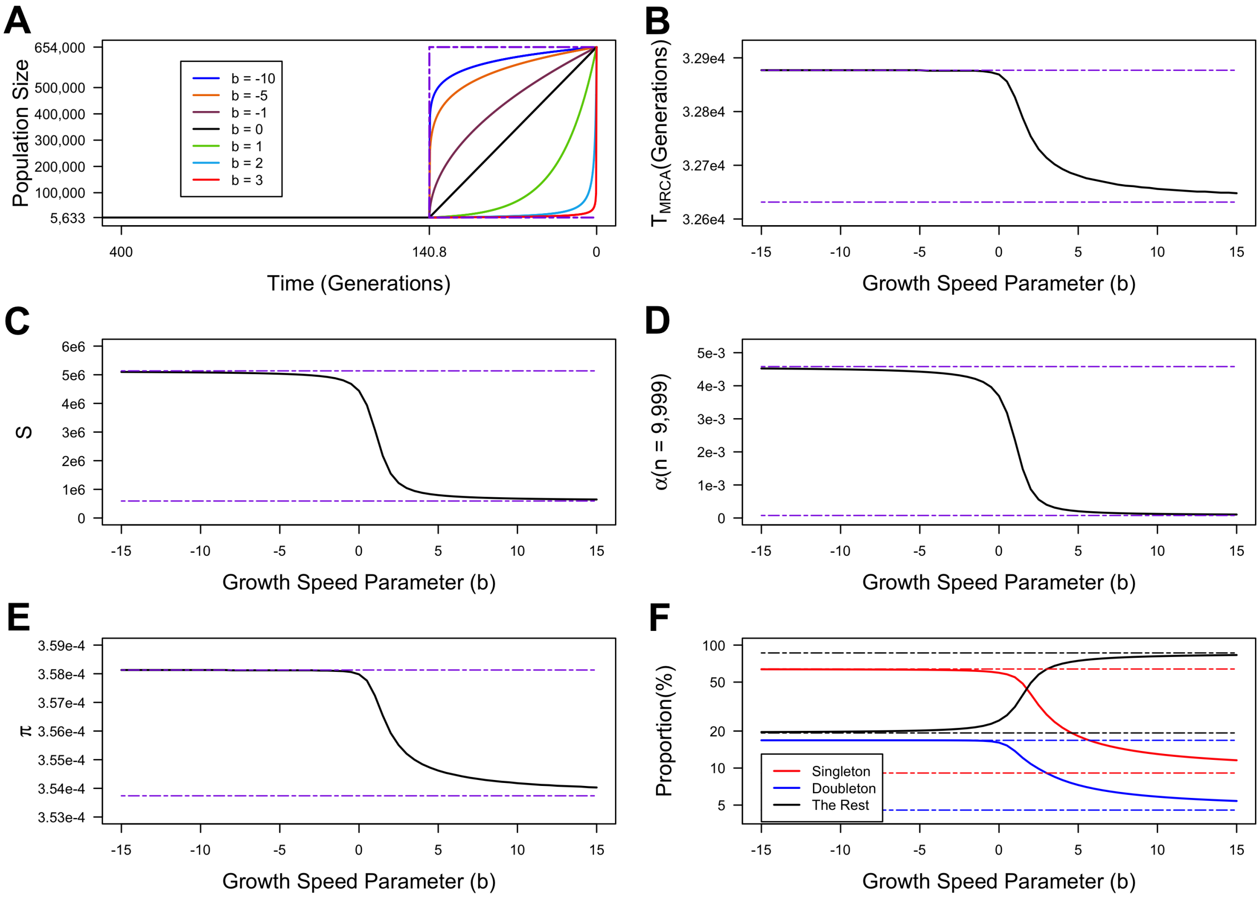

Figure S3. Expected values of summary statistics generated under demographic models with a wide range of the growth speed parameter (b). The time values and population size values are kept the same as shown in Figure 2(A). The growth speed parameter (b) of the recent epoch takes values from -15 to 15. The sample size is 10,000 individuals. The mutation rate per site per generation $\mu_0 = 1.2 \times 10^{-8}$. We assumed a total of 2×10^8 sites , thus the locus-based mutation rate $\mu = 2.4$ (same for Figure 2). (A) The demonstration of the demographic models for several values of b. To better exhibit the difference between different values of b, only the most recent 400 generations are shown. The two dotted purple lines show the constant-size model fixed at 5,633 (corresponding to $b \to \infty$) and an instant-increase model with a sudden change from 5,633 to 654,000 at 140.8 generations ago (corresponding to $b \to -\infty$). (B)-(E) The expected value of T_{MRCA}, S, α at $n =$ 9,999 and π respectively for b varying from -15 to 15. The two dotted purple lines correspond to the expected values for the scenarios shown by the dotted purple lines in (A). (F) The expected proportion of singletons (red), doubletons (blue) and the sum of the rest entries of the SFS for b varying from -15 to 15. The dotted lines show expected singletons (red), doubletons (blue) and the rest (black) of the SFS for the scenarios shown by the dotted purple lines in (A).

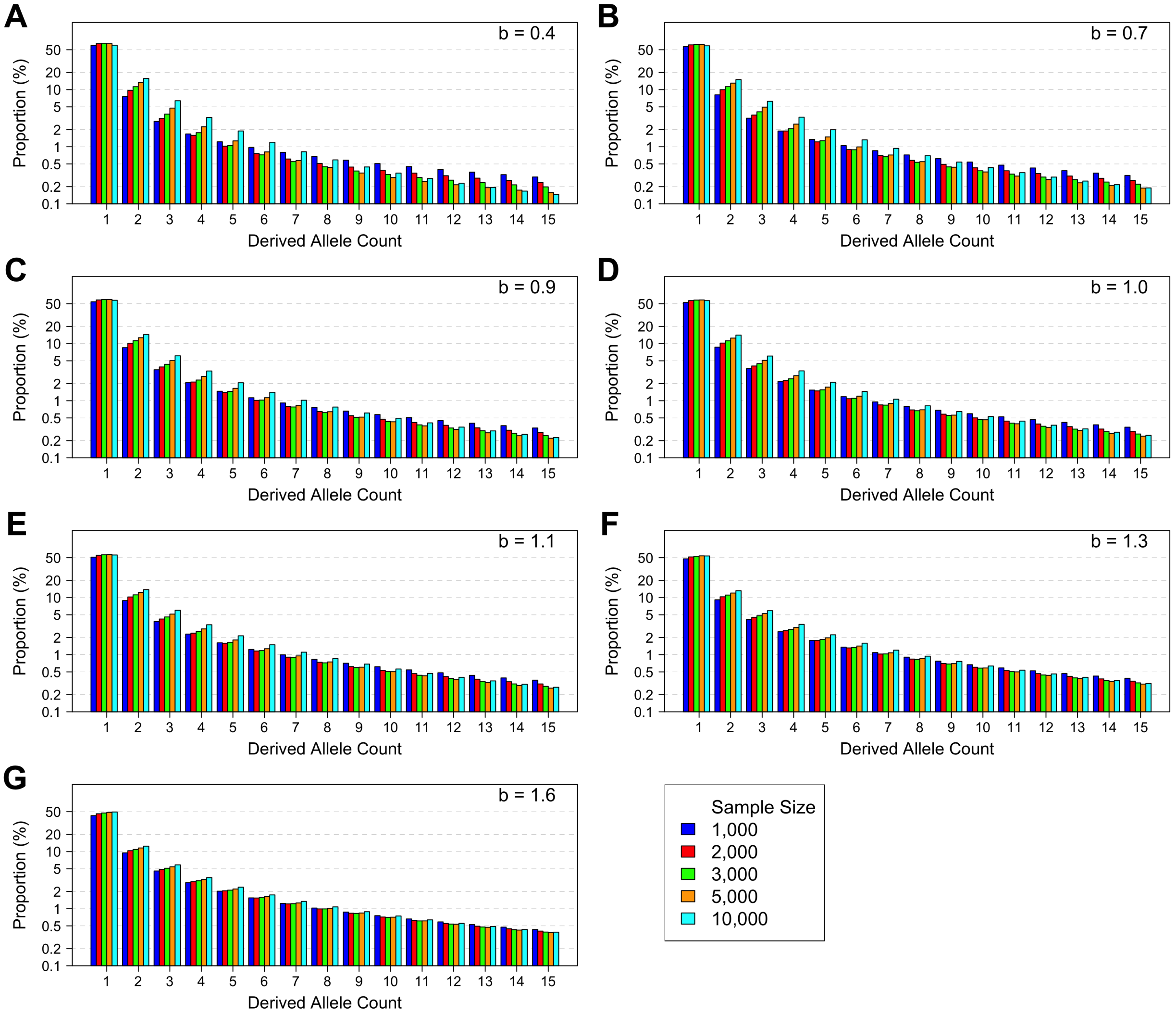

Figure S4. The first 15 entries of the site frequency spectra for the simulation scenarios described in the second section of Results. The inference results are shown in Figure 3. (A)-(G): corresponding to $b = 0.4$, $b = 0.7$, $b = 0.9$, $b = 1.0$, $b = 1.1$, $b = 1.3$ and $b = 1.6$ respectively for the recent generalized growth epoch, with sample size of 1,000 diploids (blue), 2,000 diploids (red), 3,000 diploids (green), 5,000 diploids (orange) and 10,000 diploids (cyan).

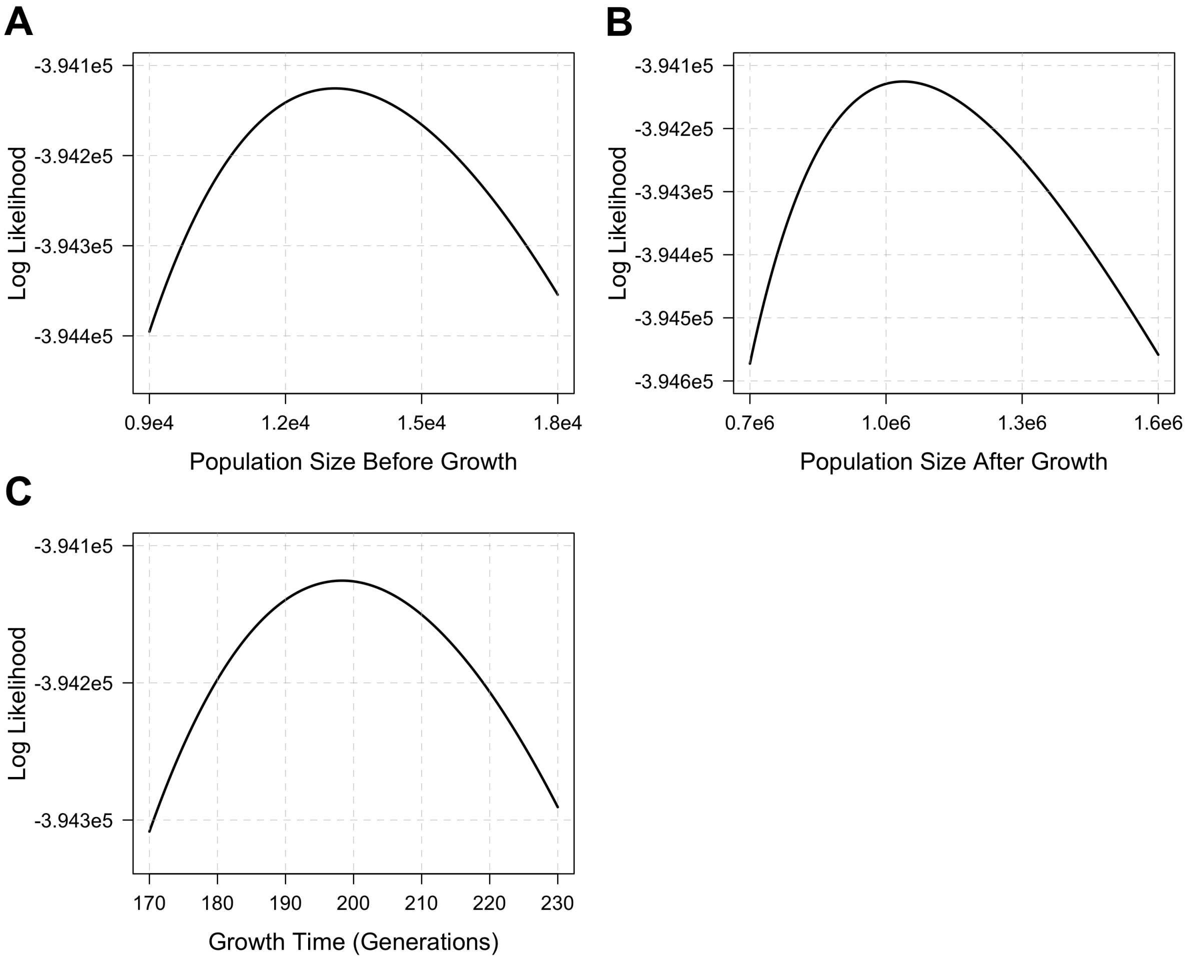

Figure S5. The one-dimensional log likelihood surface around the best estimates of the ESP synonymous data using exponential growth model. (A) varying population size before growth while keeping all other parameters at corresponding best estimates; (B) varying population size after growth only; (C) varying growth time only.

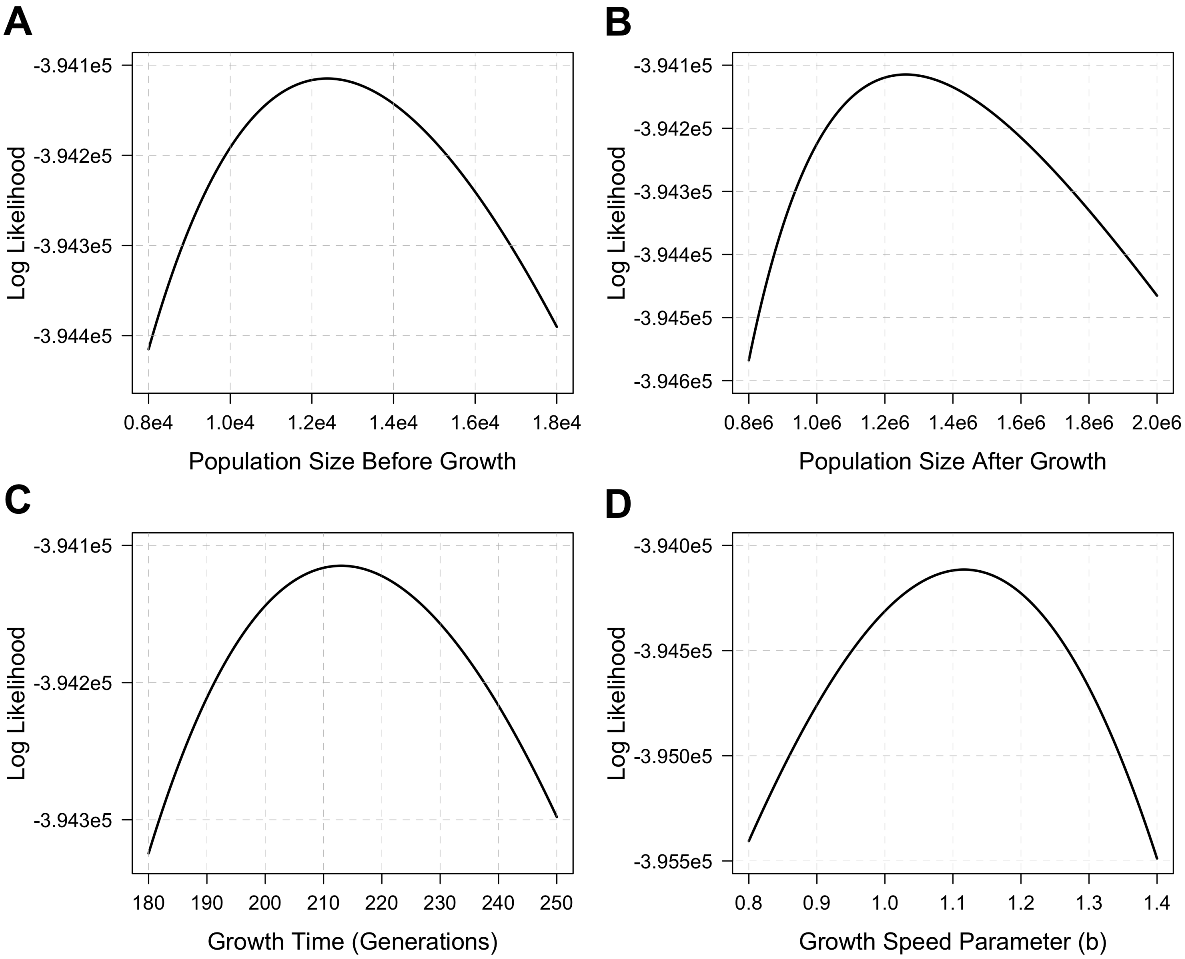

Figure S6. The one-dimensional log likelihood surface around the best estimates of the ESP synonymous data using generalized growth model. (A) varying population size before growth while keeping all other parameters at corresponding best estimates; (B) varying population size after growth only; (C) varying growth time only; (D) varying growth speed parameter only.

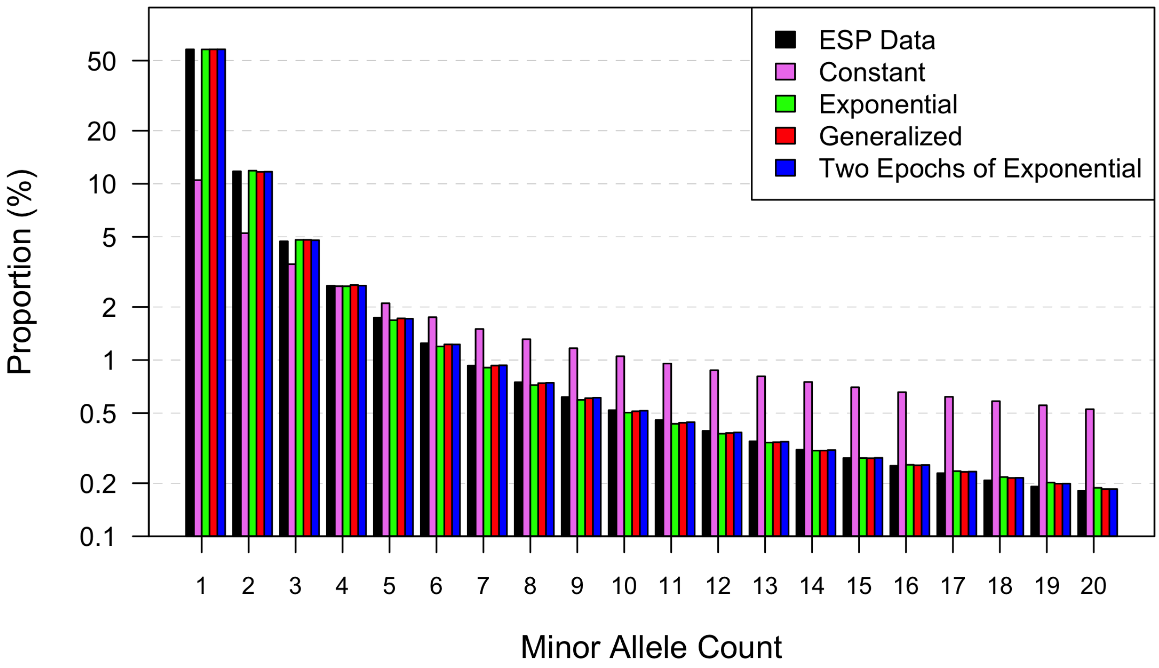

Figure S7. The first 20 entries of the site frequency spectra for ESP data and the inferred demographic models assuming the ancient demography in Gazave *et al.* (2014). The SFS from the ESP data, the exponential model, the generalized growth model and the two-epoch exponential model are shown in black, green, red and blue respectively. For comparison purposes, we also included the SFS from a base model, which has a constant population size throughout history (in pink).

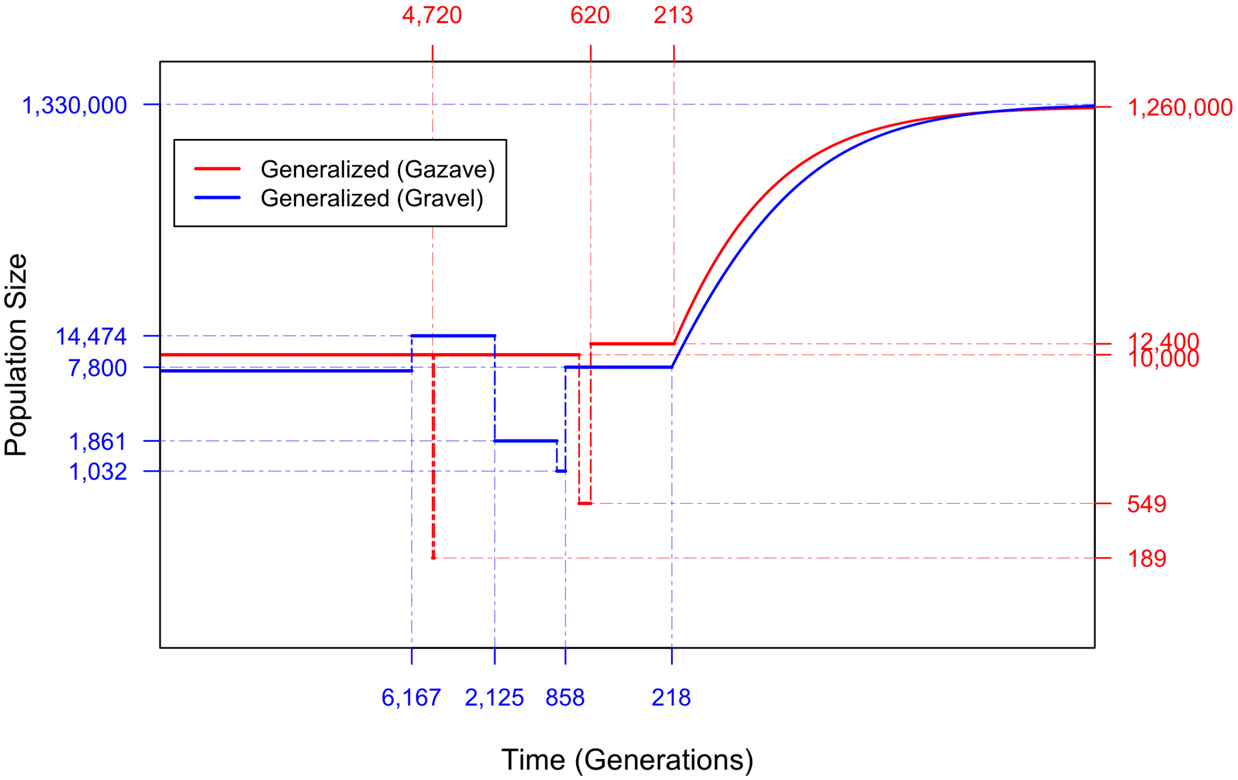

Figure S8. The best-fit generalized models for ESP data assuming the ancient demography in Gazave *et al.* (2014) (red) and in Gravel *et al.* (2011) (blue). The demographic history was fixed before 620 generations ago for Gravel model and 858 generations ago for Gravel model. Both x-axis and y-axis are on log scale.

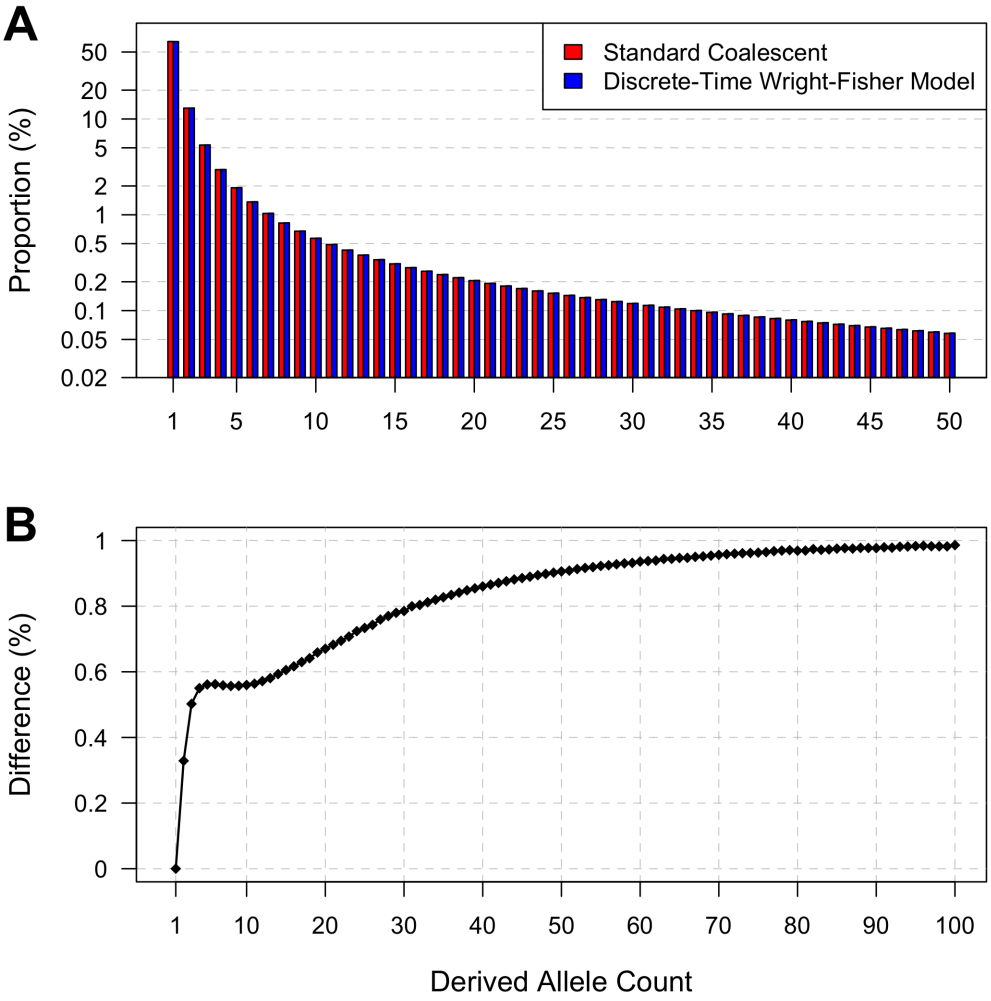

Figure S9. Effects of multi-merger and simultaneous-merger events on the SFS. The underlying demographic model is the best-fit generalized model using the ancient history in Gazave *et al.* (2014). The sample size is 3,870 diploid individuals. (A) The 100-entry *partially normalized* SFS under Kingman's coalescent and under discrete-time Wright-Fisher model. (B) The percentage difference of entry-to-singleton ratio between Kingman's coalescent and discrete-time Wright-Fisher model for the first 100 entries.

Table S1. Comparison of summary statistics computed by EGGS and estimated by FTEC simulation. Only 2,000 loci (1,000 bp-long each) were simulated for the demographic models shown in Figure 2(A). Presented are (i) the total number of segregating sites (S) across all 2,000 loci (1,000 bp-long each), (ii) the mean pairwise difference between chromosomes per base pair (π), and (iii) the burden of private mutation (α) as the percentage of heterozygous variants in one individual that are monomorphic in the rest of the sample of 999 individuals.

		Values of b		
		0.5	1.0	1.5
$S(10^{-4})$	EGGS	10.06	9.70	7.72
	FTEC	10.06	8.96	7.73
$\pi(10^{-4})$	EGGS	3.58	3.57	3.57
	FTEC	3.53	3.49	3.56
$\alpha(10^{-3})$	EGGS	7.56	5.97	4.18
	FTEC	7.66	6.00	4.24

Table S2. Demographic inference results using ESP data for a model with two recent epochs of exponential growth. Shown are point estimates and 95% confident intervals (in parenthesis) for the following parameters of the inferred epoch: population size before growth (N_2), population size after the more ancient phase of exponential growth (N_1), population size after the recent phase of exponential growth (N_0), time when the ancient phase of exponential growth started (T_2, in generations), time when the recent phase of exponential growth started (T_1, in generations).

$N_2(10^4)$	$N_1(10^4)$	$N_0(10^6)$	T_2	T_1
1.22 (1.17 $\sim$ 1.32)	4.71 (3.03 $\sim$ 54.09)	1.12 (1.07 $\sim$ 2.09)	219 (95 $\sim$ 334)	135 (25 $\sim$ 157)

Table S3. Goodness of fit between the SFS from inferred models and ESP data. We show the p-value from χ^2 goodness of fit test and KL divergence between the SFS from the ESP data and that from the constant population size model, the inferred exponential model, the generalized model and the two-epoch exponential model. The assumed ancient history (Gazave model or Gravel model) is indicated in parenthesis. The constant population size model is included here for comparison purposes.

Model	p-value from χ^2 test	KL divergence
Constant	0	0.84
Exponential (Gazave)	0.97	1.64×10^{-4}
Generalized (Gazave)	1	1.15×10^{-4}
Two-Epoch Exponential (Gazave)	1	1.09×10^{-4}
Exponential (Gravel)	1.59×10^{-6}	4.12×10^{-4}
Generalized (Gravel)	1	1.15×10^{-4}

File S1

1 Detailed description of genetic summary statistics

1.1 Total number of segregating sites (S)

Suppose we have n sequences (chromosomes), this quantity stands for the number of sites in which the sequences have different genotypes. Namely, if all sequences have a common genotype for a site, this site is not considered as a segregating site.

1.2 Time to the most recent common ancestor (T_{MRCA})

This statistic is the time taken for all of the samples at present to coalesce to the same ancestor.

1.3 Site frequency spectrum (SFS)

Suppose we have n sequences sampled at present, the full SFS $\boldsymbol{\xi}$ has $(n-1)$ entries $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_{n-1})$, where ξ_i records the fraction of segregating sites that have i derived alleles and $(n-i)$ ancestral alleles. When we don't have information about the ancestral allele, the folded SFS $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_{\lfloor \frac{n}{2} \rfloor})$ is used, where η_i records the fraction of segregating sites that have i minor alleles and $(n-i)$ major alleles. By definition, $\eta_i = \dfrac{\xi_i + \xi_{n-i}}{1 + \delta(i, n-i)}$.

1.4 Average pairwise difference per site (π)

Suppose we have n sequences sampled at present. We compare every two different sequences (thus there are $\binom{n}{2}$ pairs), count the number of differences between each pair, calculate the average of the total differences and normalize the average difference by the total number of sites, or total length of loci L. This quantity has the following relationship with the SFS and S:

$$\pi = \frac{S}{L\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i = \frac{S}{L\binom{n}{2}} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} i(n-i)\eta_i.$$

1.5 Burden of private mutations (α)

Suppose we have n diploid individuals sequenced (thus there are $2n$ sequences). α stands for the proportion of heterozygous positions in a newly sequenced $(n+1)^{\mathrm{th}}$ individual that are novel. Namely, all of the previous n individuals have the same genotype at such a site, but this newly sequenced individual have a different genotype.

2 More detailed explanation of the growth speed parameter b_k

When $r_k \neq 0$, the growth speed is controlled by the parameter b_k. With the same value of r_k, $N_{k,\mathrm{f}}$ and $(T_k - T_{k-1})$, if $b_k > 1$, the model will reach a $N_{k,\mathrm{i}}$ larger than that of an exponential model. As a result, it is considered to be faster than exponential or super-exponential. Similarly, if $b_k < 1$, the model will reach a $N_{k,\mathrm{i}}$ smaller than that of an exponential model and thus is considered to be slower than exponential or sub-exponential.

To illustrate the above facts, we give an example in Figure S1(A). The growth epoch starts 200 generations ago with a population size of 10,000. The value of growth rate $g = \frac{d}{dT} \log N(T)$ is fixed at 0.35% such that when exponential growth model is used, the population size after growth is 20,000, which is a 2-fold growth. The values of b are chosen to be 0.9, 1 and 1.1. When $b = 1.1$, the population size after growth is 67,730, larger than 20,000 when exponential growth is considered. Similarly, when $b = 0.9$, the population size after growth 13,129, smaller than 20,000.

If we fix $N_{k,\mathrm{i}}$, $N_{k,\mathrm{f}}$ and $(T_k - T_{k-1})$, as is mostly considered in this study, taking different values of b will cause the growth pattern to be different. When $b > 1$, the growth will show an accelerating pattern compared with exponential growth; while when $b < 1$, the growth will show a decelerating pattern. To illustrate this point, consider the models shown in Figure S1(B). The growth epoch is from 200 generations ago to present and the population sizes before and after growth are fixed at 10,000 and 100,000 respectively. The values of b are chosen to be 0.3, 1 and 1.7. For the exponential model, the growth rate 1.15% is constant throughout the epoch. For $b = 1.7$, the growth rate (0.52%) is smaller than that of the exponential growth (1.15%) at the onset time of 200 generations ago. The growth keeps accelerating as time approaches present. At $t = 0$, the growth rate for $b = 1.7$ (2.87%) is larger than that of the exponential (1.15%). For $b = 0.3$, the pattern is opposite. The instantaneous growth rate (2.87%) is larger than that of the exponential growth (115.13) at 200 generations ago. The growth keeps decelerating as time approaches present. At $t = 0$, the instantaneous growth rate for $b = 0.3$ (0.57%) is smaller than that of the exponential (1.15%).

3 Quantities A_j^p, V_j^p and $W_{i,j}^p$

For computing $\mathbb{E}[T_{\mathrm{MRCA}}^p]$, the quantities A_j^p can be calculated by (Polanski *et al.* 2003; Tavare 1984; Takahata and Nei 1985)

$$A_j^p = \frac{(-1)^j (2j-1) p_{[j]}}{p^{(j)}},$$

where $p_{[j]}$ is the falling factorial function, $p_{[j]} = p(p-1)\cdots(p-j+1)$, and $p^{(j)}$ is the rising factorial function, $p^{(j)} = p(p+1)\cdots(p+j-1)$.

For computing $\mathbb{E}[\boldsymbol{\xi}^p]$, the quantities V_j^p can be calculated by (Polanski and Kimmel 2003)

$$V_j^p = (2j-1)\frac{p!(p-1)!}{(p+j-1)!(p-j)!}[1+(-1)^j],$$

and $W_{i,j}^p$ are constants given by the following recursive relationships (Polanski and Kimmel 2003):

$$W_{i,2}^p = \frac{6}{p+1}; W_{i,3}^p = \frac{30(p-2i)}{(p+1)(p+2)}; W_{i,j+2}^p = -\frac{(1+j)(3+2j)(p-j)}{j(2j-1)(p+j+1)}W_{i,j}^p + \frac{(3+2j)(p-2i)}{j(p+j+1)}W_{i,j+1}^p.$$

4 Expressions of r_k

For the generalized growth models considered in this study, any epoch k is determined by the starting population size $N_{k,\mathrm{i}}$, the ending population size $N_{k,\mathrm{f}}$, the duration of the epoch $(T_k - T_{k-1})$ and the growth speed parameter b_k. After determining the epoch, the dependent parameter $r_k = r_k(N_{k,\mathrm{i}}, N_{k,\mathrm{f}}, b_k, T_k - T_{k-1})$ is calculated by

$$r_k = \begin{cases} \dfrac{N_{k,\mathrm{i}}^{1-b_k} - N_{k,\mathrm{f}}^{1-b_k}}{T_k - T_{k-1}}, & b_k \neq 1 \\ \dfrac{\log N_{k,\mathrm{i}} - \log N_{k,\mathrm{f}}}{T_k - T_{k-1}}, & b_k = 1 \end{cases}.$$

5 Expressions of $\Lambda(T)$ for evaluating ϕ_j^k

For convenient purposes, define $\lambda_k(T) = \int_{T_{k-1}}^{T} d\sigma/\mathcal{N}(\sigma)$, where $\mathcal{N}(\sigma) = 2N(\sigma)$ and $T_{k-1} \leq T \leq T_k$, then $\Lambda(T) = \Lambda(T_{k-1}) + \lambda_k(T)$. For generalized models, the solution for $\lambda_k(T)$ is

$$\lambda_k(T) = \begin{cases} \dfrac{T - T_{k-1}}{\mathcal{N}_{k,\mathrm{i}}}, & r_k = 0 \\ \dfrac{\log \mathcal{N}_{k,\mathrm{i}} - \log \mathcal{N}(T)}{r_k}, & b_k = 0, r_k \neq 0 \\ \dfrac{\mathcal{N}(T)^{-b_k} - \mathcal{N}_{k,\mathrm{i}}^{-b_k}}{b_k r_k}, & b_k \neq 0, r_k \neq 0 \end{cases}.$$

Notice that the third expression above is also true for exponential growth/decline ($b_k = 1$ and $r_k \neq 0$).

6 Evaluation of ϕ_j^k for non-linear non-exponential generalized decline epochs

Generally, under arbitrary population size function $N(T)$, the quantity

$$\phi_j^k = \mathrm{e}^{-\binom{j}{2}\Lambda(T_{k-1})} \int_{T_{k-1}}^{T_k} \mathrm{e}^{-\binom{j}{2}\lambda_k(T)}\, dT,$$

where $\Lambda(T) = \int_0^T d\sigma/\mathcal{N}(\sigma)$, $\lambda_k(T) = \int_{T_{k-1}}^T d\sigma/\mathcal{N}(\sigma)$ and $\mathcal{N}(\sigma) = 2N(\sigma)$.

For generalized decline epochs ($r_k < 0$ and $b_k \notin \{0, 1\}$), in which case we didn't find feasible closed-form expression for evaluating ϕ_j^k, this quantity can be expressed in the following way:

$$\phi_j^k = \frac{\mathrm{e}^{-\binom{j}{2}\Lambda(T_{k-1})}}{\binom{j}{2}} \int_0^{\frac{\binom{j}{2}}{b_k r_k}\left(\mathcal{N}_{k,\mathrm{f}}^{-b_k} - \mathcal{N}_{k,\mathrm{i}}^{-b_k}\right)} \left(\frac{b_k r_k}{\binom{j}{2}} y + \mathcal{N}_{k,\mathrm{i}}^{-b_k}\right)^{-\frac{1}{b_k}} \mathrm{e}^{-y}\, dy.$$

The integral $\int_0^{\frac{\binom{j}{2}}{b_k r_k}\left(\mathcal{N}_{k,\mathrm{f}}^{-b_k} - \mathcal{N}_{k,\mathrm{i}}^{-b_k}\right)} \left(\frac{b_k r_k}{\binom{j}{2}} y + \mathcal{N}_{k,\mathrm{i}}^{-b_k}\right)^{-\frac{1}{b_k}} \mathrm{e}^{-y}\, dy$ is in the form of $\int_0^d (ax + b)^c e^{-x}\, dx$ where a, b, c, d are constants. We numerically evaluate this integral by Gauss-Legendre quadrature (Kahaner *et al.* 1988). The basic idea of Gauss-Legendre quadrature is to approximate the integrated function $f(x) = (ax + b)^c e^{-x}$ by a polynomial function of degree n, and evaluate $f(x)$ at n different points in the range $[0, d]$. The error term is $\frac{d^{(2n+1)} n!^4}{(2n+1)(2n)!^3} f^{2n}(\xi)$ (Kahaner *et al.* 1988), where $0 < \xi < d$ and $f^{(2n)}$ is the $(2n)^{\text{th}}$ derivative of f with respect to x. We choose the polynomial degree n to be 512 in this work.

7 Libraries used/adapted in this study

For the computation of functions $\mathcal{U}(b, x)$ and $\mathcal{M}(b, x)$, we adapted the C++ codes for the evaluation of confluent hypergeometric functions from GSL scientific library (Galassi *et al.*). In addition, we used the library from the link `http://www.holoborodko.com/pavel/numerical-methods/numerical-integration/`, which is provided by Pavel Holoborodko for Gauss-Legendre quadrature. The authors are grateful to the providers of these libraries, which are essential in the implementation of the `EGGS` software.

8 Details of simulation parameters in the second section of Results

When simulating the sequences, we used mutation rate $\mu = 1.2 \times 10^{-8}$ per base pair per generation (Kong *et al.* 2012) and recombination rate $\rho = 1.0 \times 10^{-8}$ per base pair per generation. To determine the amount of data for simulation, we used the number of exomes given in Tennessen *et* *al.* (2012), which is about 2,500 and assumed that each exome has 20,000 base pairs on average. To stress more the effect of linkage disequilibrium (LD) between the alleles in each exome, we decreased the number of independent loci to 1,000 and increased the length of each locus to 50,000, while keeping the total number of base pairs the same. To reduce noise in the simulated data and increase computation speed, we only kept the first 100 entries of the folded SFS and calculated the aggregate sum of the rest entries, such that there are 101 entries in total.

9 Details of bootstrapping

We used 200 bootstraps to obtain 95% confidence interval of the inferred parameters. For simulation studies, we randomly choose 1,000 loci from the simulated 1,000 independent loci with replacement in each bootstrap. For inference based on ESP data (Tennessen *et al.* 2012; Fu *et al.* 2013), we split the sequences into 500kb regions based on SNP positions, which resulted in 882 different regions, similar to the number of loci in simulation studies. In the same manner, we then chose 882 regions with replacement for each bootstrap.

10 Subsampling approach

For ESP data, the successful genotype counts vary across different segregating sites. We applied the subsampling approach similarly considered in Gazave *et al.* (2014) and Gao and Keinan (2014). For a site with n successful genotype counts, suppose there are j minor alleles and $(n-j)$ major alleles, the probability that it is of x minor alleles when subsampled to m chromosomes is

$$\mathbb{P}[x \leftarrow m] = \frac{\binom{j}{x}\binom{n-j}{m-x}}{\binom{n}{m}} + \frac{\binom{j}{m-x}\binom{n-j}{x}}{\binom{n}{m}}$$

where $x = 0, 1, 2, \cdots, \lfloor \frac{m}{2} \rfloor$. In this work, we choose m (the number of chromosomes to subsample to) to be 7,740, which is 90% of the total number of chromosomes (8,600).

11 Composite log likelihood

In order to determine the fitness of a model Θ to the observed folded allele frequency counts $\boldsymbol{\mathcal{C}}$, we compute the log likelihood of the model according to

$$\mathbb{L}[\Theta] = \log \mathbb{P}[\boldsymbol{\mathcal{C}} \mid \Theta] = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \mathcal{C}_i \log \mathbb{E}[\eta_i \mid \Theta],$$

where $\mathbb{E}[\boldsymbol{\eta}|\Theta] = \left(\mathbb{E}\left[\eta_1 \mid \Theta\right], \mathbb{E}\left[\eta_2 \mid \Theta\right], \cdots, \mathbb{E}\left[\eta_{\lfloor \frac{n}{2} \rfloor} \mid \Theta\right]\right)$ is the expected folded SFS given model Θ. In this work, we considered SFS binning from the 101st entry to reduce the noise in later parts of the SFS: $\mathbb{E}\left[\widetilde{\boldsymbol{\eta}} \mid \Theta\right] = \left(\mathbb{E}\left[\eta_1 \mid \Theta\right], \mathbb{E}[\eta_2 \mid \Theta], \cdots, \mathbb{E}\left[\eta_{100} \mid \Theta\right], \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \mathbb{E}\left[\eta_i \mid \Theta\right]\right)$, and correspondingly the binned allele frequency counts from the data $\widetilde{\boldsymbol{\mathcal{C}}} = \left(\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_{100}, \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \mathcal{C}_i\right)$. The log likelihood after binning is computed as

$$\mathbb{L}[\Theta] = \log \mathbb{P}[\widetilde{\boldsymbol{\mathcal{C}}} \mid \Theta] = \sum_{i=1}^{101} \widetilde{\mathcal{C}}_i \log \mathbb{E}[\widetilde{\eta}_i \mid \Theta].$$

12 Goodness of fit measures

In order to test how well a model SFS fits the observed data, we performed χ^2 goodness of fit test. In specific, if the observed allele frequency counts is $\boldsymbol{\mathcal{C}} = (\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_d)$ (which indicates that the total number of observed segregating sites is $|\boldsymbol{\mathcal{C}}|$) and the SFS under the model is $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_d)$, then the statistic

$$\chi^2 = \sum_{i=1}^{d} \frac{(\mathcal{C}_i - |\boldsymbol{\mathcal{C}}|\xi_i)^2}{|\boldsymbol{\mathcal{C}}|\xi_i}.$$

The degree of freedom is $(d-1)$, where d is the dimension of the vector $\boldsymbol{\mathcal{C}}$. A p-value > 0.05 means we fail to reject the null hypothesis that the observed data SFS is consistent with the model SFS.

We also used another measure, Kullback-Leibler divergence or KL divergence (Kullback and Leibler 1951), which provides a single number to relatively compare the goodness of fit between different models:

$$\mathcal{D}_{\mathrm{KL}}\left(\frac{\boldsymbol{\mathcal{C}}}{|\boldsymbol{\mathcal{C}}|} \,\middle\|\, \boldsymbol{\xi}\right) = \sum_{i=1}^{d} \frac{\mathcal{C}_i}{|\boldsymbol{\mathcal{C}}|} \left(\log \frac{\mathcal{C}_i}{|\boldsymbol{\mathcal{C}}|} - \log \xi_i\right).$$

A smaller $\mathcal{D}_{\mathrm{KL}}$ means a higher consistency between the observed and the model. The advantage of KL divergence over log likelihood is that KL divergence is a normalized measure unaffected by the total number of observed segregating sites $|\boldsymbol{\mathcal{C}}|$.

The p-values from χ^2 goodness of fit test and the KL divergence between the observed ESP data and the SFS from each of the inferred models are shown in Table S3.

13 Potential effect of multi-merger and simultaneous-merger events on the SFS

As sample size increases, the probability of multi-merger and simultaneous-merger events will rise, which violates the assumptions of Kingman's coalescent and might affect the SFS (Bhaskar and Song 2014). To test this effect, we used the discrete-time Wright-Fisher (DTWF) model software

(Bhaskar and Song 2014) to compute the SFS under the generalized Gazave *et al.* model with a sample size of 7,740. To shorten the computation time, we used a hybrid of DTWF and Kingman's coalescent with a time cutoff of $t_c = 212$ generations. However, it is still computationally burdensome to evaluate all 7,739 entries of the *unnormalized* SFS $\boldsymbol{\Xi}^{\text{DTWF}} = \left(\Xi_1^{\text{DTWF}}, \Xi_2^{\text{DTWF}}, \cdots, \Xi_{7739}^{\text{DTWF}}\right)$, which is needed to compute the *normalized* SFS $\boldsymbol{\xi}^{\text{DTWF}} = \dfrac{\boldsymbol{\Xi}^{\text{DTWF}}}{\left|\boldsymbol{\Xi}^{\text{DTWF}}\right|} = \left(\xi_1^{\text{DTWF}}, \xi_2^{\text{DTWF}}, \cdots, \xi_{7739}^{\text{DTWF}}\right)$ as is used in the inference work. We instead only evaluated the first 100 entries of $\boldsymbol{\Xi}^{\text{DTWF}}$, $\left(\Xi_1^{\text{DTWF}}, \Xi_2^{\text{DTWF}}, \cdots, \Xi_{100}^{\text{DTWF}}\right)$.

We first compared the *partially normalized* SFS under DTWF model $\boldsymbol{\xi}_{\text{partial}}^{\text{DTWF}} = \dfrac{1}{\sum_{i=1}^{100} \Xi_i} \left(\Xi_1^{\text{DTWF}}, \Xi_2^{\text{DTWF}}, \cdots, \Xi_{100}^{\text{DTWF}}\right)$ with the *partially normalized* SFS under Kingman's coalescent $\boldsymbol{\xi}_{\text{partial}}^{\text{Kingman}} = \dfrac{1}{\sum_{i=1}^{100} \xi_i^{\text{Kingman}}} \left(\xi_1^{\text{Kingman}}, \Xi_2^{\text{Kingman}}, \cdots, \Xi_{100}^{\text{Kingman}}\right)$, which was computed by `EGGS`. The two *partially normalized* SFS are very similar (Figure S9(A)). We next compared the ratio of any entry to singletons under DTWF model and Kingman's coalescent,

$$\rho_i^{\text{DTWF}} = \frac{\Xi_i^{\text{DTWF}}}{\Xi_1^{\text{DTWF}}}; \rho_i^{\text{Kingman}} = \frac{\xi_i^{\text{Kingman}}}{\xi_1^{\text{Kingman}}},$$

where $i = 1, 2, \cdots, 100$ and we calculated the relative error,

$$\epsilon(i) = \frac{\rho_i^{\text{DTWF}} - \rho_i^{\text{Kingman}}}{\rho_i^{\text{Kingman}}} \times 100\%,$$

where $i = 1, 2, \cdots, 100$. The relative error is always less than 1% for the first 100 entries and asymptotically increases to 1% (Figure S9(B)). We then used 1% as the relative error for the rest of the SFS entries to predict the *full normalized* SFS under DTWF model. This predicted folded SFS is very similar to the folded SFS under Kingman's coalescent (KL divergence = 6.14×10^{-6}) and fits almost equally well to the data (KL divergence between the predicted SFS and ESP data = 1.24×10^{-4}; p-value from χ^2 goodness of fit test = 1).

References

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

Fu, W., T.D. O'Connor, G. Jun, H.M. Kang, G. Abecasis *et al.*, 2013 Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493(7431):216-220.

Galassi, M., *et al.* GNU Scientific Library Reference Manual (3rd Ed.), ISBN 0954612078.

Gao, F., and A. Keinan, 2014 High burden of private mutations due to explosive human population growth and purifying selection. *BMC Genomics* 15 Suppl 4:S3.

Gazave, E., L. Ma, D. Chang, A. Coventry, F. Gao *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.

Kahaner, D., C.B. Moler, S. Nash, and G.E. Forsythe, 1988 Numerical methods and software. Englewood Cliffs, N.J.: Prentice Hall.

Kong, A., M.L. Frigge, G. Masson, S. Besenbacher, P. Sulem *et al.*, 2012 Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412):471-475.

Kullback, S., and R. A. Leibler, 1951 On information and sufficiency. *Ann Math Stat* 22:79-86.

Polanski, A., A. Bobrowski, and M. Kimmel, 2003 A note on distributions of times to coalescence, under time-dependent population size. *Theoretical Population Biology* 63(1):33-40.

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(1):427-436.

Takahata, N., and M. Nei, 1985 Gene Genealogy and Variance of Interpopulational Nucleotide Differences. *Genetics* 110(2):325-344.

Tavare, S., 1984 Line-of-Descent and Genealogical Processes, and Their Applications in Population-Genetics Models. *Theoretical Population Biology* 26(2):119-164.

Tennessen, J.A., A.W. Bigham, T.D. O'Connor, W. Fu, E.E. Kenny *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64-69.