Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models

Feng Gao*† and Alon Keinan*†

October 9, 2015

*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853
Short Running Title: Inference of generalized growth

Key Words: coalescent, generalized models, population growth, human demographic history, software

†Corresponding Author:
Alon Keinan
Department of Biological Statistics and Computational Biology
Cornell University
Ithaca, NY 14853
ak735@cornell.edu

Feng Gao
Department of Biological Statistics and Computational Biology
Cornell University
Ithaca, NY 14853
fg237@cornell.edu
ABSTRACT

The site frequency spectrum (SFS) and other genetic summary statistics are at the heart of many population genetics 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 excess amount of extremely rare variants unexplained. 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 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 decent sample sizes facilitate accurate inference, e.g. a sample of 3,000 individuals with the amount of data expected from exome sequencing allows observing and accurately estimating growth with speed deviating by 10% or more 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.
INTRODUCTION

Summary statistics of genetic variation play a vital role in population genetics 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 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; Tennesen et al. 2012; Gazave et al. 2014). However, the above studies assumed that the growth is exponential. The observation of huge amount of extremely rare, previously unknown variants in several sequencing studies with large sample sizes (Nelson et al. 2012; Tennesen et al. 2012; Fu et al. 2013) and the recent explosive growth in census population size suggests that human population might have experienced a recent super-exponential growth, i.e. growth with speed faster than exponential (Coventry et al. 2010; Keinan and Clark 2012; Reppell et al. 2012; Reppell et al. 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; Reppell et al. 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: (1) the time to the most recent common ancestor ($T_{\text{MRCA}}$), (2) the total number of segregating sites ($S$), (3) the site frequency spectrum (SFS), (4) the average pairwise difference between chromosomes per site ($\pi$), and (5) the burden of private mutations, BPM ($\alpha$), 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, EGGS (Efficient computation of Generalized models’ Genetic summary Statistics), that 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 facilitates computation that is orders of magnitudes faster than that of simulations. By performing demographic inference on the SFS generated from simulated sequences, we then explored how many samples are needed for recovering parameters of a recent generalized growth epoch. Finally, we applied the software to investigate the nature of the recent growth in humans by inferring demographic models using the SFS of synonymous variants of 4,300 European individuals from the NHLBI Exome Sequencing Project (Tennessen et al. 2012; Fu et al. 2013).
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, 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 \( T_1, T_2, \ldots, T_L \) \((T_0 = 0 < T_1 < T_2 < \cdots < T_L < T_{L+1} = \infty)\), 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 \( N(T_{k-1} \leq T < T_k) \) can be described by the following differential equation regarding time \( T \) (Reppell et al. 2012; Reppell et al. 2014):

\[
\frac{dN}{dT} = -r_k N^{b_k},
\]

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 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 5 epochs is provided in Figure 1, with more detailed explanation and illustrations in File S1 and Figure S1.
The solution to Equation (1) is (Reppell et al. 2012; Reppell et al. 2014)

\[
N(T) = \begin{cases} 
  (N_{k,i}^{-b_k} - r_k (T - T_{k-1}) (1 - b_k))^{-1/b_k}, & b_k \neq 1 \\
  N_{k,i} e^{-r_k (T - T_{k-1})}, & b_k = 1
\end{cases}
\] (2)

where \( N_{k,i} \) is the initial population size of the \( k \)th epoch. Each epoch \( k \) is defined by 4 parameters: the starting population size \( N_{k,i} \), the ending population size \( N_{k,f} \), 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, \( r_k = r_k (N_{k,i}, N_{k,f}, b_k, T_k - T_{k-1}) \), 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 \( N_{k+1,i} \), the starting population size of the \((k + 1)\)th epoch is not necessarily the same as \( N_{k,f} \), the ending population size of the \( k \)th epoch. Specifically, if \( N_{k+1,i} \neq N_{k,f} \), 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; Kingman 1982b), given a demographic model \( N(T) \), the expected time of the most recent common ancestor \( \mathbb{E}[T^p_{\text{MRCA}}] \) can be calculated by (Polanski and Kimmel 2003)

\[
\mathbb{E}[T^p_{\text{MRCA}}] = \sum_{j=2}^{p} A_j^p \psi_j,
\] (3)

where the superscript \( p \) is the number of chromosomes (i.e. twice the sample size for diploids), \( \psi_j \) is the expected time of the first coalescent event when there are \( j \) chromosomes at present and \( A_j^p \) 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 $N(T) = 2N(T)$. Then $\psi_j$ is expressed by the following equation:

$$
\psi_j = \int_0^\infty T \frac{1}{N(T)} e^{-\int_0^T \frac{1}{N(T)} d\sigma} dT = \int_0^\infty e^{-(\frac{1}{2})\Lambda(T)} dT,
$$

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

The expected full normalized SFS $E[\xi^p] = (E[\xi^p_1], E[\xi^p_2], \ldots, E[\xi^p_{p-1}])$ can be computed by the following set of equations (Polanski et al. 2003):

$$
E[\xi^p] = \frac{E[\ell^p_i]}{E[L^p]}, E[\ell^p_i] = \sum_{j=2}^p W^p_{i,j} \psi_j; E[L^p] = \sum_{j=2}^p V^p_j \psi_j,
$$

where $\ell^p_i$ is the length of branches in the genealogy that have $i$ descendants ($i = 1, 2, \cdots, p-1$) and $L^p = \sum_{i=1}^{p-1} \ell^p_i$ 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

$$
E[S] = \mu_0 L E[L^p],
$$

where $\mu_0$ 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 $E[\pi]$ can be calculated by

$$
E[\pi] = 2\mu_0 E[T^{p=2}_{MRCA}].
$$

The expected burden of private mutations $\alpha$ at diploid sample size of $(\frac{p}{2} - 1)$, defined as the proportion of heterozygous sites in a new diploid individual that are homozygous in the previous $(\frac{p}{2} - 1)$ individuals, $E[\alpha_{\frac{p}{2}-1}]$ can be computed by (Gao and Keinan 2014)

$$
E[\alpha_{\frac{p}{2}-1}] = \frac{2}{p[1 + \delta(1, p - 1)]} \frac{E[\ell^p_1] + E[\ell^p_{p-1}]}{E[\ell^p_1]},
$$

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 \( \psi_j \), the expected time of the first coalescent event when there are \( j \) chromosomes at present. Previous studies give explicit expressions of \( \psi_j \) for a demographic model constructed by exponential and constant-size epochs \( \text{[Polanski et al. 2003; Bhaskar et al. 2013]} \). In this study, we give a comprehensive set of formulas for generalized models introduced above. Define \( \phi_j^k := \int_{T_{k-1}}^{T_k} e^{-(\frac{j}{2})A(T)} \, dT \), then \( \psi_j = \sum_{k=0}^{L+1} \phi_j^k \), where \( (L + 1) \) is the total number of epochs. The quantity \( \phi_j^k \) can be computed by the following set of equations:

1. If \( r_k = 0 \) or \( b_k = 0 \), \( r_k \neq 0 \):
   \[
   \phi_j^k = \begin{cases} 
   \frac{1}{(\frac{j}{2})} \left[ e^{-(\frac{j}{2})A(T_k)} N_{k,f} \log N_{k,f} - e^{-(\frac{j}{2})A(T_{k-1})} N_{k,i} \log N_{k,i} \right], & r_k + (\frac{j}{2}) = 0 \\
   \frac{1}{r_k + (\frac{j}{2})} \left[ e^{-(\frac{j}{2})A(T_{k-1})} N_{k,i} - e^{-(\frac{j}{2})A(T_k)} N_{k,f} \right], & r_k + (\frac{j}{2}) \neq 0 
   \end{cases} \; ; \quad (9)
   \]

2. If \( b_k > 0, r_k > 0 \) or \( b_k = 1, r_k < 0 \):
   \[
   \phi_j^k = \frac{1}{(\frac{j}{2})} \left[ N_{k,i} \mathcal{U} \left( 2 - \frac{1}{b_k}, \frac{(\frac{j}{2})}{b_k r_k} N_{k,f} \right) e^{-(\frac{j}{2})A(T_{k-1})} - N_{k,f} \mathcal{U} \left( 2 - \frac{1}{b_k}, \frac{(\frac{j}{2})}{b_k r_k} N_{k,f} \right) e^{-(\frac{j}{2})A(T_k)} \right]; \quad (10)
   \]

3. If \( b_k < 0, r_k > 0 \):
   \[
   \phi_j^k = \frac{1}{(\frac{j}{2})} \left[ N_{k,i} \mathcal{U} \left( 2 - \frac{1}{b_k}, \frac{(\frac{j}{2})}{b_k r_k} N_{k,f} \right) e^{-(\frac{j}{2})A(T_k)} - N_{k,i} \mathcal{U} \left( 2 - \frac{1}{b_k}, \frac{(\frac{j}{2})}{b_k r_k} N_{k,f} \right) e^{-(\frac{j}{2})A(T_{k-1})} \right]; \quad (11)
   \]

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

\[
U(1, x) = x e^x \int_1^\infty \frac{e^{-t}}{t} dt = x e^x E_1(x),
\]

(12)

where \(E_1(x)\) is the Exponential Integral (Gradshteyn 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 \(\phi_j^k\) 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 (Efficient computation of Generalized models’ Genetic summary Statistics). The source code and compiled programs for Linux and Mac OS platforms are publicly available from our website (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, Gazave et al. (2014) model inferred that European population had a constant effective population size of 10,000 (diploid) individuals before 4,720 generations ago, and went through the ancient bottleneck between 4,720 and 4,620 generations ago with a population size of 189. The population size then recovered to 10,000 diploids until 720 generations ago, from which time
the recent bottleneck started with a size of 549. At 620 generations ago, the population size recovered to 5,633 individuals. The recent growth epoch started from 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 European population had an ancient effective population size of 7,300 diploid individuals until 6,167 generations ago, when the population size expanded to 14,474 individuals. The first bottleneck took place 2,125 generations ago, with the population size reducing to 1,861 individuals. This first bottleneck lasted until 958 generations ago, at which time a second bottleneck took place with a decreased population size of 1,032. We assumed 24 years per generation (Scally and Durbin 2012) to translate the year-based time presented in the original model. For compatibility with 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 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 dataset. To determine the fitness of a model $N(T)$ to the observed data, we calculated the composite log likelihood by

$$L[N] = \log E[\xi | N] = C \cdot E[\xi | N],$$

(13)

where $C$ is a vector of the observed folded allele frequency counts and $E[\xi | N]$ is the computed folded SFS under demographic model $N(T)$. More detailed description can be found in File
To search for the maximum likelihood point over the parameter space, we applied the ECM method (Meng and Rubin 1993), which was previously used in the demographic inference study by Excoffier et al. (2013). 100 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 contains $l$ 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) contains deep sequencing of 4,300 individuals of European ancestry. An important feature of these data is the high sequencing coverage, which allows capturing very rare variants accurately. These variants constitutes 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 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 less than or equal to 20, or with successful genotype counts smaller than 7,740 (90%), and subsampled the remaining 233,134 SNVs to 7,740 alleles, which is equivalent to 3,870 diploid individuals (File S1).
RESULTS

COMPARISON WITH SIMULATED RESULTS BY FTEC

To validate that the expressions provided in the Methods and Materials section 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 2(A). 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 1,000 diploid individuals (2,000 chromosomes). For FTEC simulation, we used a mutation rate of $1.2 \times 10^{-8}$ per base pair per generation (e.g., Kong et al. 2012) and simulated 200,000 independent loci, each of 1,000 base pairs.

The comparison of the SFS, $S$ (across all 200,000 loci), $\pi$ and $\alpha$ numerically computed by EGGS to that simulated by FTEC is shown in Figure 2(B)-(E). For each demographic model illustrated in Figure 2(A), 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 2(A), EGGS takes less than a second to generate the results, while it takes about 5 hours for FTEC to simulate the sequences, due to the large number of independent loci required for accurate estimation (performed in Ubuntu system with Intel Xeon CPU @ 2.67GHz). For instance, when 2,000 independent loci are simulated, which still takes about 3 minutes, 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
EVALUATING INference OF GENERATED 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. \textbf{Bhaskar and Song (2014)} has shown 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 2(A)). However, the SFS is estimated in practice from 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 \textbf{(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 parameter and other parameters of the demographic model. 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 2(A). Starting 620 generations ago, the simulated model included a constant population size of 10,000 until 200 generations ago, when the population starts a generalized growth epoch till 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 ($b = 1.0$) in order 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 1,000, 2,000, 3,000, 5,000 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 infer 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 \( N_i \); 3) the ending population size after growth \( N_f \); 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 1,000 diploids is used for inference, the inference performs badly for all parameters with large confidence intervals (Figure 3). However, the confidence interval always includes the true simulated value. A sample size of 2,000 already exhibits acceptable performance except when the growth speed becomes large (\( b = 1.3 \) and 1.6). Larger sample sizes of 5,000 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 (1.1) growth has taken place. These observations suggest that a sample size of at least 3,000 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 Exome Sequencing Project (ESP) data \cite{Tennessen2012,Fu2013}. We applied our inference 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 \cite{Gazave2014} model illustrated in Figure 2(A). 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 were: for both models (1) population size before growth ($N_f$), (2) population size after growth ($N_i$), (3) growth onset time ($T$), which is equivalent to the duration of growth; 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(AB).

Although \cite{Gazave2014} model assumed a different ancient history before the recent growth epoch than that assumed in \cite{Tennessen2012}, using ESP data and assuming exponential growth, the inferred growth epoch is generally consistent with that obtained in their study (Figure 4(AB) and Table 1). Our study infers that recent growth started 198 (95% CI: 195-202) generations ago with an effective population size of $\sim$13,100 (12,600-13,600) and continued at a rate of 2.2% (2.15%-2.26%) per generation (Table 1), while \cite{Tennessen2012} estimated that recent growth had an initial population size of $\sim$9,500 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 $\chi^2$ likelihood-ratio test with 1 degree of freedom). It estimates that 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 exponential speed of $b = 1$ ($p$-value $\ll 10^{-12}$ using 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 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 (Methods and Materials). We repeated inference of the same parameters using the same ESP data. The inferred parameters for exponential growth are also 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 Gazave et al. (2014) model (\( p \)-value = \( 1.59 \times 10^{-6} \) from \( \chi^2 \) goodness of fit test between exponential Gravel model and ESP data, which is 0.97 for exponential Gazave model; see File S1 and Table S3). By applying a generalized growth epoch to Gravel et al. (2011) model, the inferred parameters are generally in line with those from generalized Gazave model, 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, generalized Gravel model fits the data almost equally well as generalized Gazave model, which is significantly better than the exponential model (\( p \)-value \( \ll 10^{-12} \) by \( \chi^2 \) likelihood-ratio test; also see Table S3). As with generalized Gazave model, the inferred growth speed parameter from 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 one-tailed \( z \)-test; Figure 4(CD)).

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 Gazave et al. (2014) model illustrated in Figure 2(A). 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 $\chi^2$ likelihood-ratio test with 2 degrees of freedom), but is a worse model than the *generalized* growth model (based on Bayesian Information Criterion, $\text{BIC}_{\text{two-epoch-exponential}} - \text{BIC}_{\text{generalized}} = 6.1$). However, this model exhibits some of the same accelerating pattern 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 Modern Era (Keinan and Clark 2012).

**DISCUSSION**

In this study, we provide the 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 ($T_{\text{MRCA}}$), the total number of segregating sites ($S$), the site frequency spectrum (SFS), the average pairwise difference between chromosomes per site ($\pi$) and the burden of private mutation ($\alpha$). 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 generalized models by extending 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 of the first coalescent event ($\psi_j$). In these cases, fast and accurate numerical integration methods, such as Gauss-Legendre quadrature used in this work, can be applied to evaluate $\psi_j$. 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 non-negligible. 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 discrete-time Wright-Fisher (DTWF) model (Bhaskar et al. 2014) under the inferred demographic history in generalized Gazave model with a sample size of 3,870 diploids (File S1). We observed that the SFS from 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 on the SFS generated from the synonymous variants of 4,300 individuals of NHLBI ESP dataset (Tennessen et al. 2012, Fu et al. 2013), we found that generalized growth model shows a better fit to the observed data than the
exponential growth model that has been used by all previous demographic modeling studies ($p$-value = $3.85 \times 10^{-6}$). We also find that the European population experiences 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 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 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 is 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 (Taylor 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 models can be estimated with generalized growth models.
ACKNOWLEDGEMENTS

The authors would like to thank Leo 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


Taylor, P. N., E. Porcu, S. Chew, P. J. Campbell, M. Traglia, S. J. Brown,


FIGURES AND TABLES

Figure 1. Illustration of an example of a generalized demographic model as introduced in the first section of Methods and Materials. This model consists of 5 epochs (starting from present): (1) Faster-than-exponential \((b > 1)\) growth (looking 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 growth 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 \(T_4\); (5) Constant population size (a special case of generalized growth when \(r = 0\)) at \(N_{5,i} = N_{5,f}\) starting from \(T_4\) and 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.
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)](http://example.com), 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)](http://example.com), 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 (1,000 bp-long each) (C), the expected pairwise difference between chromosomes per base pair (D) and 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 (E) computed numerically in EGGS (dark bars) and simulated by FTEC (light bars) for the demographic models shown in (A): $b = 0.5$, blue; $b = 1.0$, black; $b = 1.5$, red, with a sample size of 1,000 individuals (2,000 chromosomes). The y-axis in (B) is on log scale.
Figure 3. Inference results on simulated data with a recent generalized growth epoch. The model parameters are as follows: Growth starts 200 generations before 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 (1,000, black; 2,000, red; 3,000, blue; 5,000, brown; and 10,000, green). The point estimates with 95% confidence interval for these parameters are grouped by the growth speed parameter $b$ ($x$-axis). The dashed purple 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, (D) the inferred growth.
start time. The $y$-axis in (C) is on log scale.
Figure 4. Demographic inference results based on ESP data. (A) Illustration of the effective population size (y-axis, on 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 2(A). Solid lines show the effective population size over time of each of the inferred models, with dashed line indicating estimated parameter values on the x-axis or y-axis. Only the most recent 1,000 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 y-axis on linear scale. (C) Similar to panel (A), 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). (D) A zoom-in to the most recent 240 generations of the inferred models in (C).
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. Shown are point estimates and 95% confident intervals (in parenthesis) for the following parameters of the inferred recent growth epoch when the ancient history was assumed to be the same as that in \textit{Gazave et al. (2014)} model and \textit{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.

<table>
<thead>
<tr>
<th>Ancient History</th>
<th>Growth Model</th>
<th>$N_f (10^4)$</th>
<th>$N_i (10^6)$</th>
<th>$T$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gazave Model</td>
<td>Exponential</td>
<td>1.31</td>
<td>1.04</td>
<td>198</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.26-1.36)</td>
<td>(1.00-1.07)</td>
<td>(195-202)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Generalized</td>
<td>1.24</td>
<td>1.26</td>
<td>213</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.18-1.30)</td>
<td>(1.16-1.37)</td>
<td>(206-220)</td>
<td>(1.07-1.15)</td>
</tr>
<tr>
<td>Gravel Model</td>
<td>Exponential</td>
<td>0.89</td>
<td>0.85</td>
<td>186</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.86-0.93)</td>
<td>(0.82-0.88)</td>
<td>(182-190)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Generalized</td>
<td>0.78</td>
<td>1.33</td>
<td>218</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.74-0.83)</td>
<td>(1.22-1.46)</td>
<td>(211-228)</td>
<td>(1.18-1.26)</td>
</tr>
</tbody>
</table>